

UNIVERSITY OF
BIRMINGHAM

MECHANISM AND FUNCTIONAL ANALYSIS OF CHAPERONINS OF *Mycobacterium marinum*

By

SHAHIDA RAFIQUE BUTT

A thesis submitted to the University of Birmingham for the degree of
DOCTOR OF PHILOSOPHY

Institute of Microbiology & Infection

School of Biosciences

College of Life and Environmental Sciences

University of Birmingham

July 2022

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

Mycobacterial chaperonins have been demonstrated to play an important role in immunomodulation. Both the chaperonins of *M. tuberculosis* have been revealed to be involved directly or indirectly in pathology of tuberculosis. In a study, MtbCpn60.1 was found to be essential for granulomatous inflammation of infected mice and guinea pigs. We used *M. marinum* as a model organism to study the mechanism and functions of mycobacterial chaperonins. The genome comparison of *M. marinum* and *M. tuberculosis* and the phylogenetic analysis of 50 mycobacterial species and 12 other species of actinobacteria showed that both the organisms are close genetic relatives and share ~50% homologous coding sequences in their genomes. To investigate the promoter regions of the chaperonins of *M. marinum* different lengths of the upstream regions of the chaperonin genes of *M. marinum* were cloned in pSD5B which is a promoter probe shuttle vector. The strongest promoter activity was observed in the fragment upstream to Mm*cpn60.2* followed by that of Mm*cpn10*. No promoter activity was found in the intergenic region of Mm*cpn60.1*.

RNA-seq analysis of *M. marinum* WT, Δ*cpn60.1*, and complemented strains was performed to study the transcriptome profile and the effect of deletion of cpn60.1 on the expression profile of *M. marinum*. The TPM values of the read counts of the triplicate samples of WT *M. marinum* showed that the expression level of *cpn60.2* was slightly higher than 2-fold than that of *cpn10.* The expression of *cpn10* was nearly doubled the expression of *cpn60.1* indicating the differential expression of bicistronic *cpn10-cpn60.1*. In addition, DESeq2 analysis of the RNA-seq data demonstrated that the deletion of Mm*cpn60.1* significantly differentially expressed (upregulated in KO) the sets of genes involved in three major sets of pathways. First, the highest fold enrichment pathways were related to stress response, chaperone activity and protein folding. The second highest fold enrichment pathways were related to cell wall, cell division, and peptidoglycan biosynthesis whereas the third highest fold enrichment pathways were related to polyketide synthase pathways. The results indicated that although not essential for cell survival but Mm*cpn60.1* is directly or indirectly involved in regulation of key mechanisms of the cell.

The complementation analysis of the chaperonins of *M. marinum* demonstrated that MmCpn60.2 was able to complement the loss of GroEL in *E. coli* in either combination with MmCpn10 or GroES whereas MmCpn60.1 failed show complementation phenotype in all attempts. The results supported the previous findings that the Cpn60.2 plays the role of housekeeping chaperonin. The “sequence optimized”, RBS distance optimized, and double phosphomimetic mutant of MmCpn60.1 also failed to demonstrate the chaperone function in *E. coli*. MmCpn60.1 might perform chaperone function in its natural cell environment as indicated in DESeq2 analysis by significant upregulation of chaperone genes in KO mutant strain. Further investigation of the role of Cpn60.1 in peptidoglycan and cell wall lipids biogenesis would be advantageous in understanding its role in the pathogenesis of tuberculosis.

Dedication

This dissertation is dedicated to my mother, my father, my sisters, my brothers, my nephews, and my nieces. Thank you all for your support, inspiration, encouragement, and constant love which have sustained me throughout my life. Lots of prayers for my sister “Musarrat Rafique” in heaven.

Acknowledgment

Dr Peter Lund has been my ideal PhD supervisor, mentor, and teacher. His immense assistance, guidance, encouragement, and support with a perfect blend of kindness, affection, and patience assisted me immensely throughout my PhD. This research project would not have been completed without his support and motivation. I would like to express my bundles of thanks to Dr Peter Lund. He not only guided me throughout my research but also helped me with my scientific writing skills. During my PhD, he also gave me intellectual freedom, supported me in shining my presentation skills and engaged me in scientific ideas. I am extremely grateful to you for your availability any time I needed guidance and support. You are not only an understanding and great supervisor but also a kind-hearted, generous, and good-natured human.

I would also like to express my huge thanks to Dr CM Santosh Kumar for his assistance and help throughout my PhD. He not only helped me in learning lots of new laboratory skills but also contributed to discussions of this research project. Thank you for your kindness and scientific plus moral support. A Special thanks to Dr Mathew Milner who assisted me with all the bioinformatics analysis of this project. Immense thanks to Dr Hrishiraj Sen who was the first person from Pete Lund group I met with. Additionally, I would like to thank my second supervisor Dr Apoorva Bhatt and my internal assessor Dr Luke Alderwick for assessing my work during assessment periods and for providing valuable feedback.

I want to extend my huge thanks to Dr Fatima Alattar, Dr Robert L. Marshall, Manpreet Kaur, Dr Jack Bryant, Gabriela Boelter, Dana Alfawaz, Jessica Gray, Max Alexander, Emily Goodall, Georgia Isom, Samantha Mckeand, Kara Staunton, Rachel Chandler, Emma Sheehan, Deema Alodaini, Nina perry, Gabriel Wong, Doaa Ghulam, Emilie Eamens, Anis Dadou, Alexander Brooks, Amina Aziz, Zahra Ahmed, and Bakul piplani. You all are nice and kind people who made T101 a lovely and dynamic laboratory.

The most important thanks to the Darwin Trust of Edinburgh for funding my PhD and providing me this life changing and marvellous opportunity. Thanks to the administrator of the Darwin Trust “Heather Hall” for her assistance. Huge thanks to Sir Kenneth Murray for his vital initiative of the trust.

I would also like to thank my Pakistani friends at Birmingham: Saad Ehsan, Arslan Mahmood, Faiza Salman, Muhammad Waqas Ali, Sana Ali, and Sidra. Thank you all for your love, moral support, encouragement, and friendship. Thanks to my supportive and kind friends at Jarratt Hall especially those who were there during COVID lockdown, Daria Nikiforova, Ronnie Wu, and Calvin Menezes. I cannot forget the support of Zainab Jalalpuria, Aminath Latheef, and other flatmates at Cadnam Hall who provided me moral support in the hardest and saddest moment of my sister's loss.

Thanks to the most important members of my life, my mother, my father, my kind sisters, my brothers, my lovely nephews, my beautiful nieces, my brothers-in-law, and my sisters-in-law who always trusted and supported me. Thanks to my mother whose love and backing made me confident and motivated. Huge thanks to my father for always trusting in me blindly. His trust and encouragement are the biggest assets of my life. Thank you all for making the best family. Lots of duas and love for my passed away sister Musarrat Rafique. She had a generous and kind heart. You will always be missed and remembered in my prayers. Special thanks to my sisters Tahira Rafique, Farhat Rafique, and Zainab Rafique for their love and trust. You have always been supporting me in difficult times. I am blessed to have lovely and kind sisters like you. My PhD would not be possible without the encouragement of my parents and my sisters. May Allah Almighty bless you all in this world and hereafter.

I cannot forget to thank my all teachers at school, college and at the Quaid-i-Azam university, Islamabad. Thanks to Dr Siddique for establishing the best science school in my hometown where there was no science school for the children before my age group. Thanks to Sir Salman Haider and Madam Bushra Kouser for being great and inspiring teachers in my BSc. And huge thanks to my teachers of Microbiology Department at the Quaid-i-Azam University, Dr. Abdul Hameed, Dr Samiullah Khan, Dr Malik Yousafzai, Dr Aamer Ali Shah, Dr Muhammad Imran, Dr Naeem Ali, Dr Asif Jamal, Dr Rabaab Zahra, Dr Rani Faryal, Dr Fariha Hasan, Dr Safia Ahmed, Dr Najma Ayub, Dr Javed Dasti, and Dr ishtiaq Ali. May Allah Almighty bless you all.

I would like to thank again to everyone who made this thesis possible. Thanks to the supportive staff of the School of Biosciences, University of Birmingham. Those whose names are mentioned and whose names are left, I pray for you all. God bless you all.

Table of contents

1 INTRODUCTION 1

1.1 TUBERCULOSIS 2
1.1.1 M. TUBERCULOSIS INFECTION 5
1.1.2 DETERMINANTS OF GRANULOMA FORMATION - SURVIVAL AND PROLIFERATION OF *M. TUBERCULOSIS* IN HOST 8
1.2 ZEBRAFISH - *M. MARINUM* MODEL 13
1.2.1 ZEBRAFISH (DANIO RERIO) – A BETTER HOST TO STUDY TUBERCULOSIS 13
1.2.2 *M. MARINUM* – A CLOSE GENETIC RELATIVE OF *M. TUBERCULOSIS* 15
1.3 MOLECULAR CHAPERONES AND CO-CHAPERONES 16
1.3.1 DEFINITION OF MOLECULAR CHAPERONES 16
1.3.2 IMPORTANCE OF MOLECULAR CHAPERONES TO THE CELL 17
1.3.3 CLASSES OF MOLECULAR CHAPERONES 20
1.3.4 EXAMPLES OF MOLECULAR CHAPERONES IN *M. MARINUM* AND *M. TUBERCULOSIS* 22
1.3.5 ROLE OF DIFFERENT CLASSES OF MOLECULAR CHAPERONES IN PROTEIN HOMEOSTASIS 23
1.3.6 DNAK/DNAJ/GRPE PROTEINS CHAPERONE COMPLEX 25
1.3.7 CHAPERONINS – GROEL/GROES COMPLEX 27
1.3.7.1 Structure of GroEL and GroES 28
1.3.7.2 GroES-GroEL protein folding mechanism 30
1.4 MULTIPLE CHAPERONINS 35
1.4.1 NOMENCLATURE OF CHAPERONINS 35
1.4.2 CHAPERONINS OF MYCOBACTERIA 36
1.4.3 OLIGOMERIC STATE 37
1.4.4 PHOSPHORYLATION 37
1.4.5 EXPRESSION OF CPN PROTEINS IN MYCOBACTERIA 38
1.4.6 ROLE IN IMMUNOMODULATION 38
1.5 AIMS AND OBJECTIVES 40

2 MATERIALS AND METHODS 42

2.1 BACTERIAL STRAINS 43
2.2 PLASMIDS 44
2.3 PRIMERS 50
2.4 GENERAL MICROBIOLOGY TECHNIQUES 57
2.4.1 GROWTH MEDIA 57
2.4.1.1 LB Broth and LB agar 57
2.4.1.2 Difco 7H9 and 7H10 57
2.4.1.3 Super Optimal Broth (SOB) 57
2.4.1.4 SOC media 57
2.4.2 MEDIA SUPPLEMENTS 58
2.4.3 GROWTH CONDITIONS 58
2.4.3.1 *E. coli* growth conditions 58
2.4.3.2 Mycobacteria growth conditions 59
2.4.4 GLYCEROL STOCKS 59
2.4.5 GENOMIC DNA EXTRACTION 59
2.4.6 AGAROSE GEL ELECTROPHORESIS 60
2.4.7 PLASMID PURIFICATION 61
2.4.7.1 Alkaline lysis method 61
2.4.7.2 Plasmid purification using QIAGEN Plasmid kits 62

2.4.8 PREPARATION OF CHEMICALLY COMPETENT CELLS 62
2.4.8.1 Calcium chloride method 63
2.4.8.2 Inoue competent cells preparation method 64
2.4.9 TRANSFORMATION 65
2.4.10 PREPARATION OF ELECTROCOMPETENT CELLS 65
2.4.10.1 Electroporation 66
2.4.11 RNA EXTRACTION 66
2.4.11.1 RNA quantification and RIN number calculation 67
2.5 QUALITATIVE AND QUANTITATIVE ASSAYS 67
2.5.1 X-GAL QUALITATIVE ASSAY/ BLUE WHITE SCREENING 67
2.5.2 B-GALACTOSIDASE ACTIVITY ASSAY (LACZ ASSAY) 68
2.5.3 COMPLEMENTATION ASSAY 69
2.6 GENERAL MOLECULAR BIOLOGY TECHNIQUES 69
2.6.1 SANGER SEQUENCING 69
2.6.2 PLASMID CLONING 70
2.6.2.1 Cloning by restriction digestion 70
2.6.2.2 Cloning of PCR fragment 71
2.6.2.3 Cloning of Oligos 72
2.6.3 SITE DIRECTED MUTAGENESIS (SDM) 73
2.7 PROTEIN ANALYSIS TECHNIQUES 74
2.7.1 SODIUM DODECYL SULFATE-POLYACRYLAMIDE GEL ELECTROPHORESIS (SDS-PAGE) 74
2.7.2 NATIVE-PAGE 76
2.7.3 WESTERN BLOTTING 77
2.7.4 PROTEIN PURIFICATION 78
2.7.4.1 Nickle-nitrilotriacetic acid Ni-NTA column protein purification 78
2.7.4.2 Immunoprecipitation 80
2.8 BIOINFORMATICS ANALYSIS 82
2.8.1 GENOME BROWSERS 82
2.8.2 PHYLOGENETIC TREE CONSTRUCTION 82
2.8.3 ROARY ANALYSIS 82
2.8.4 RNA-SEQ ANALYSIS 82

3 GENOME COMPARISON AND PHYLOGENETIC RELATIONSHIP 83

3.1 INTRODUCTION 84
3.2 ANALYSIS PLAN 87
3.3 RESULTS 88
3.3.1 WHOLE GENOME COMPARISON OF *M. MARINUM* AND *M. TUBERCULOSIS* 88
3.3.1.1 Hypothetical proteins and PE/PPE family proteins 88
3.3.1.2 Other specific enzymes 89
3.3.1.3 Transport proteins and other secretion system proteins 90
3.3.1.4 Lipoproteins, transcriptional regulatory proteins, toxins/anti-toxins, and virulence proteins 90
3.3.1.5 Plasmid pMMRN 91
3.3.1.6 Construction of genome maps of 1218R and H37Rv 91
3.3.1.7 Comparison of genomes using the Artemis comparison tool (ACT) and MAUVE alignment 91
3.3.1.8 Roary analysis 96
3.3.2 PHYLOGENETIC ANALYSIS 99
3.3.3 MULTIPLE SEQUENCE ALIGNMENT OF CHAPERONINS (CPN60) 104
3.4 DISCUSSION 111

4 ANALYSIS OF PROMOTER ACTIVITY OF REGIONS UPSTREAM OF *CPN* GENES IN *M. MARINUM* 116

4.1 INTRODUCTION 117
4.1.1 EXPRESSION PATTERN OF CHAPERONIN GENES OF *M. TUBERCULOSIS* 119
4.1.2 EXPRESSION PATTERN OF CHAPERONIN GENES OF *M. SMEGMATIS* 120
4.1.3 RNASEQ AND RIBOSOME PROFILING DATA OF *CPN* GENES OF *M. TUBERCULOSIS* AND *M. SMEGMATIS* 121
4.1.4 REGULATION MECHANISM OF HEAT-SHOCK GENES IN BACTERIA 127
4.1.5 REGULATION MECHANISM OF HEAT-SHOCK GENES IN *MYCOBACTERIA* 130
4.2 BIOINFORMATICS ANALYSIS 132
4.2.1 BIOINFORMATICS ANALYSIS OF PROMOTER REGIONS OF *CPN* GENES 134
4.2.2 ANALYSIS OF MRNA SECONDARY STRUCTURES IN THE *CPN10-CPN60.1* INTERGENIC REGION 146
4.2.3 BIOINFORMATICS ANALYSIS OF HRCA REPRESSOR 150
4.3 EXPERIMENTAL PLAN 153
4.4 RESULTS 156
4.4.1 CLONING OF FRAGMENTS INTO PSD5B 156
4.4.1.1 Direct cloning of PCR fragments into pSD5B 156
4.4.1.2 Two step cloning strategy using pUC18 156
4.4.2 QUALITATIVE X-GAL ASSAY 160
4.4.2.1 Qualitative assay with *M. marinum* strains 160
4.4.2.2 Qualitative assay with *M. smegmatis* and *E. coli* 160
4.4.3 QUANTITATIVE B-GALACTOSIDASE ACTIVITY ASSAY 162
4.4.3.1 Quantitative β-galactosidase assay in *M. marinum* W.T and KO strains at normal growth conditions 163
4.4.3.2 Quantitative β-galactosidase assay in *M. marinum* W.T and KO strains at temperatures 5°C lower and higher than the optimum temperature i.e., 30°C. 165
4.4.3.3 Quantitative β-galactosidase assay in *M. marinum* W.T and KO strains at pH stress 166
4.4.3.4 Quantitative β-galactosidase assay in *M. smegmatis* 169
4.5 DISCUSSION 170

5 TRANSCRIPTOME ANALYSIS OF *M. MARINUM* BASED ON RNA-SEQ 172

5.1 INTRODUCTION 173
5.1.1 COMPARISON OF GENE EXPRESSION ANALYSIS METHODS 173
5.1.2 NEXT GENERATION SEQUENCING 175
5.1.3 RNA-SEQ WORKFLOW 181
5.1.4 BIOINFORMATICS ANALYSIS OF RNA-SEQ DATA 183
5.1.4.1 Quality control 183
5.1.4.2 Data cleaning/Pre-processing 183
5.1.4.3 Alignment of reads to the reference genome 184
5.1.4.4 Quantification of reads 184
5.1.4.5 Baseline expression 185
5.1.4.6 Differential gene expression 186
5.2 EXPERIMENTAL PLAN 189
5.3 RESULTS 193
5.3.1 RNA QUANTIFICATION AND INTEGRITY CHECK 193
5.3.2 RNA SEQUENCING – PERFORMED BY THE GENEWIZ COMPANY 194
5.3.3 BIOINFORMATICS ANALYSIS OF RNA-SEQ DATA 195
5.3.3.1 Quality control using FastQC 195
5.3.3.2 Alignment/Mapping of the sequencing reads to the reference sequences 197
5.3.3.3 SAM to BAM conversion, sorting, and indexing 198

5.3.3.4 Quantification of reads 199
5.3.3.5 Differential gene expression analysis 200
5.3.4 ANALYSIS OF DESEQ2 RESULTS, INTERPRETATION, AND VISUALIZATION OF DATA 203
5.3.4.1 Scatter plots of normalized counts 203
5.3.4.2 Heatmap of normalized counts 206
5.3.4.3 Scatter plots of log2FoldChange values of pairwise comparison 208
5.3.4.4 Differentially expressed genes 211
5.3.4.5 Venn diagrams – common and exclusive up and down regulated genes. 213
5.3.4.6 Scatter plots of differentially expressed common genes 216
5.3.4.7 Volcano plots 220
5.3.5 GENE ONTOLOGY ENRICHMENT ANALYSIS 224
5.3.5.1 Pathway analysis of common genes UKvWT_DCvKO 224
5.3.5.1.1 Analysis using *M. marinum* M homologous genes – log2FC cut off 1, padj value < 0.01 224
5.3.5.1.2 Analysis using *M. marinum* M homologous genes – log2FC between 1 – 0.585, padj value < 0.01 230
5.3.5.1.3 Role of the genes associated with cell wall biogenesis identified in pathway analysis in their respective pathways. 235
5.3.5.2 Pathway analysis of common genes DKvWT_UCvKO 240
5.3.6 PE/PPE FAMILY GENES, HYPOTHETICAL PROTEINS, AND DUF DOMAIN CONTAINING PROTEINS 243
5.3.7 RELATIVE EXPRESSION OF CHAPERONIN GENES IN WT SAMPLES 243
5.4 DISCUSSION 245

6 EXPRESSION AND COMPLEMENTATION ANALYSIS OF *M. MARINUM* CHAPERONINS IN *E. COLI* 247

6.1 INTRODUCTION 248
6.2 EXPERIMENTAL PLAN 251
6.3 RESULTS 258
6.3.1 CONSTRUCTION OF PLASMIDS 258
6.3.1.1 Construction of pTrc-*groES*-Mm*cpn60.1* and pTrc-*groES*-Mm*cpn60.2* 258
6.3.1.2 Construction of pTrc-Mm*cpn10*-Mm*cpn60.1* and pTrc-Mm*cpn10*-Mm*cpn60.2* 259
6.3.2 COMPLEMENTATION ASSAY IN *E. COLI* MGM100 260
6.3.3 EXPRESSION ANALYSIS USING SDS-PAGE 263
6.3.4 COMPLEMENTATION ASSAY IN *E. COLI* AI90 265
6.3.5 CODON OPTIMIZATION OF MM*CPN60.1* USING CODON USAGE ANALYSIS 267
6.3.5.1 Expression analysis of optimized Mm*cpn60.1* and Mm*cpn10* 269
6.3.6 OPTIMIZATION OF RBS DISTANCE FROM THE START CODON OF MM*CPN60.1* 270
6.3.6.1 Complementation and expression assay using sequence optimised and RBS optimised constructs 270
6.4 *E. COLI* GROEL WORKS AS A FUNCTIONAL CHAPERONIN IN COMBINATION WITH MMCPN10 272
6.5 DISCUSSION 275

7 FURTHER ANALYSIS OF MMCPN60.1 FUNCTION BY COMPLEMENTATION IN *E. COLI* 277

7.1 INTRODUCTION 278
7.2 EXPERIMENTAL PLAN 283
7.3 RESULTS 288
7.3.1 SERINE/THREONINE PROTEIN KINASES OF *M. MAIRNUM* 1218R 288

7.3.2 SINGLE AND DOUBLE PHOSPHOMIMETIC MUTATIONS AT S75 AND S393 POSITIONS OF MMCPN60.1-CO 289
7.3.2.1 Complementation assay 289
7.3.2.2 Expression analysis 294
7.3.3 PHOSPHOMIMETIC MUTATIONS OF SERINE AND THREONINE RESIDUES IN THE EQUATORIAL DOMAIN (TOWARDS C-TERMINAL) OF MMCPN60.1 294
7.3.4 COMPLEMENTATION ANALYSIS OF WT-CO AND DPM OF MM*CPN60.1* IN COMBINATION WITH *E. COLI GROES*. 296
7.3.5 WESTERN BLOTTING EXPERIMENT TO COMPARE THE EXPRESSION LEVELS OF MMCPN60.1-CO (W.T, DPM, AND DPM-8RBS) EXPRESSED IN COMBINATION WITH GROES AND MMCPN10 298
7.3.6 EFFECT OF ATPASE MUTATIONS ON THE COMPLEMENTATION ABILITY OF MMCPN60.1-CO PROTEINS (W.T, DPM, DPM-8RBS) IN COMBINATION WITH GROES. 299
7.3.7 EFFECT OF SWAPPING OF C-TERMINI BETWEEN MTB*CPN60.1*-DPM AND MM*CPN60.1*-DPM ON COMPLEMENTATION ABILITY OF PHOSPHOMIMETIC MUTANTS 303
7.4 MASS SPECTROMETRY ANALYSIS 310
7.4.1 PROTEIN PURIFICATION 310
7.5 LC-MS/MS ANALYSIS 310
7.6 DISCUSSION 312

8 CONCLUSION 314

List of Figures

FIGURE 1.1 THE ESTIMATED NUMBER OF REPORTED TUBERCULOSIS CASES (A) AND DEATHS (B) WORLDWIDE FROM THE YEAR 2000 – 2020. 4
FIGURE 1.2 TREATMENT RESULTS OF NEW AND RELAPSE TUBERCULOSIS CASES AND MDR-TB / RR-TB CASES WORLDWIDE FROM THE YEAR 2012 – 2019. 4
FIGURE 1.3 GRANULOMA FORMATION AND PROGRESSION IN THE LUNGS OF HUMAN FOLLOWING INHALATION OF *M. TUBERCULOSIS* CONTAMINATED AEROSOLS. 7
FIGURE 1.4 DIFFERENT TYPES OF *M. TUBERCULOSIS* INFECTION AND PERCENTAGE OF THE RISK OF EACH TYPE. 7
FIGURE 1.5 A SUMMARY OF DIFFERENT STUDIES DEMONSTRATING THE ROLE OF SOME BACTERIAL DETERMINANTS IN GRANULOMA FORMATION, THE SURVIVAL AND PROLIFERATION OF *M. TUBERCULOSIS* IN THE STRESSED ENVIRONMENT OF GRANULOMAS. 12
FIGURE 1.6 *M. MARINUM* INFECTION LEADS TO THE GRANULOMA FORMATION IN ZEBRAFISH EMBRYOS. 16
FIGURE 1.7 THE ENERGY LANDSCAPE OF PROTEIN FOLDING SHAPED BY MOLECULAR CHAPERONES. 19
FIGURE 1.8 CHAPERONE ASSISTED PROTEIN FOLDING AND PROTEIN HOMEOSTASIS. 24
FIGURE 1.9 THE DOMAIN REGIONS AND CRYSTAL STRUCTURE OF DNAK PROTEIN OF *E. COLI*. 26
FIGURE 1.10 GENOMIC ORGANIZATION OF *DNAK* , *DNAJ*, AND *GRPE* GENES IN *E. COLI* GENOME. 27
FIGURE 1.11 A GENERALIZED MODEL OF MECHANISM OF CHAPERONE ASSISTED FOLDING. 27
FIGURE 1.12 GENOMIC ORGANIZATION OF GROES-GROEL COMPLEX OF *E. COLI*. 28
FIGURE 1.13 DOMAIN REGIONS AND CRYSTAL STRUCTURE OF GROES AND GROEL. 30
FIGURE 1.14 A MODEL FOR GROES-GROEL PROTEIN FOLDING MECHANISM. FIGURE ADAPTED FROM (LUND, 2001). 32
FIGURE 1.15 CONFORMATIONAL CHANGE IN THE STRUCTURE OF GROEL COMPLEX UPON SUBSTRATE BINDING. 33
FIGURE 1.16 THE GENOMIC ORGANIZATION OF CPN60.2 (A) AND CPN60.1 (B) GENES OF *M. MARINUM* 1218R. 36
FIGURE 1.17 THE GENOMIC ORGANIZATION OF CPN60.2 (A) AND CPN60.1 (B) GENES OF *M. TUBERCULOSIS* H37RV. 36
FIGURE 3.1 MAXIMUM LIKELIHOOD TREE BASED ON 304 CORE GENES OF 157 MYCOBACTERIAL SPECIES. 85
FIGURE 3.2 GRAPHICAL REPRESENTATION OF THE GENOME MAP OF 1218R AND H37RV. 92
FIGURE 3.3 GRAPHICAL ILLUSTRATION OF M. MARINUM 1218R PLASMID PMMRN. 93
FIGURE 3.4 GENOME COMPARISON OF 1218R, H37RV, AND MC2155 USING ACT. 94
FIGURE 3.5 GENOME ALIGNMENT OF MYCOBACTERIAL GENOMES USING MAUVE MULTIPLE ALIGNMENT TOOL. 95
FIGURE 3.6 RESULTS OF THE ALIGNMENT (FIGURE 3.5) SHOWN WITHOUT LCB CONNECTION LINES AND WITH SOLID LCBS. 95
FIGURE 3.7 PANGENOME ANALYSES OF 1218R, H37RV, AND MC2155 USING ROARY 60% CUT OFF. 97
FIGURE 3.8 VENN DIAGRAM SHOWING THE RESULTS OF GENOME COMPARISON AMONG 1218R, H37RV, AND MC2155 USING ROARY ANALYSIS. 98
FIGURE 3.9 THE PHYLOGENETIC TREE BASED ON 16S RRNA CODING GENE SEQUENCE. 101
FIGURE 3.10 PHYLOGENETIC TREE BASED ON AMINO ACID SEQUENCE OF CHAPERONINS. 102
FIGURE 3.11 PHYLOGENETIC TREE BASED ON AMINO ACID SEQUENCE OF CPN60.1 AND CPN60.3. 103
FIGURE 3.12 THE PHYLOGENETIC TREE BASED ON AMINO ACID SEQUENCE OF CPN60.2. 104
FIGURE 3.13 PAIRWISE AMINO ACID IDENTITY (%) AND SIMILARITY (%) FOR HOMOLOGOUS PROTEINS. 106
FIGURE 3.14 MULTIPLE SEQUENCE ALIGNMENT OF CPN60 HOMOLOGUES OF TEN MYCOBACTERIAL SPECIES AND E. COLI GROEL. 107
FIGURE 3.15 MULTIPLE SEQUENCE ALIGNMENT OF CPN60.2 HOMOLOGUES. 109
FIGURE 3.16 MULTIPLE SEQUENCE ALIGNMENT OF CPN60.1 HOMOLOGUES. 110
FIGURE 4.1 RIBOSOME PROFILING AND RNA-SEQ GRAPHS OF *CPN10* AND *CPN60.1* OF *M. TUBERCULOSIS*. 122
FIGURE 4.2 RIBOSOME PROFILING AND RNA-SEQ GRAPHS *CPN60.2* OF *M. TUBERCULOSIS*. 123
FIGURE 4.3 RIBOSOME PROFILING AND RNA-SEQ GRAPHS OF *CPN10* AND *CPN60.1* OF *M. SMEGMATIS*. 124

FIGURE 4.4 RIBOSOME PROFILING AND RNA-SEQ GRAPHS OF CPN60.2 OF M. SMEGMATIS. 125
FIGURE 4.5 HRCA BINDING WITH THE CIRCE ELEMENT UNDER NORMAL CONDITIONS. 129
FIGURE 4.6 HRCA BINDING WITH THE CIRCE ELEMENT FOLLOWING HEAT-SHOCK. 129
FIGURE 4.7 UPSTREAM REGION OF *M. MARINUM CPN10* AND INTERGENIC REGION OF *CPN10-CPN60.1*. 135
FIGURE 4.8 UPSTREAM REGION OF *M. MARINUM CPN60.2*. 136
FIGURE 4.9 UPSTREAM REGION OF *M. MARINUM CPN10* AND INTERGENIC REGION OF *CPN10-CPN60.1* SHOWING G+C PERCENTAGE. 137
FIGURE 4.10 UPSTREAM REGION OF *M. MARINUM* CPN60.2 SHOWING G+C PERCENTAGE. 137
FIGURE 4.11 ALIGNMENT OF UPSTREAM REGION OF CPN10 GENE OF 50 SPECIES OF MYCOBACTERIA. 140
FIGURE 4.12 ALIGNMENT OF UPSTREAM REGION OF CPN10 GENE OF 12 SPECIES. 141
FIGURE 4.13 ALIGNMENT OF UPSTREAM REGION OF CPN10 GENE OF 62 SPECIES. 142
FIGURE 4.14 ALIGNMENT OF UPSTREAM REGION OF CPN60.2 GENE OF 50 SPECIES OF MYCOBACTERIA. 143
FIGURE 4.15 ALIGNMENT OF UPSTREAM REGION OF *CPN60.*2 GENE OF 12 SPECIES. 144
FIGURE 4.16 ALIGNMENT OF UPSTREAM REGION OF CPN60.2 GENE OF 62 SPECIES. 145
FIGURE 4.17 MFE STRUCTURE OF GROES-GROEL INTERGENIC REGION OF *E. COLI* MG1655. 146
FIGURE 4.18 MFE STRUCTURE ENCODING BASE-PAIR PROBABILITIES OF CPN10-CPN60.1 INTERGENIC REGION. 147
FIGURE 4.19 ALIGNMENT OF UPSTREAM REGION OF CPN60.1 GENE OF 50 SPECIES OF MYCOBACTERIA. 148
FIGURE 4.20 THE MFE STRUCTURE CREATED USING SEQUENCE ALIGNMENT. 149
FIGURE 4.21 MULTIPLE SEQUENCE ALIGNMENT OF HRCA HOMOLOGUES. 151
FIGURE 4.22 PAIRWISE AMINO ACID IDENTITY (%) AND DISTANCE OF HRCA FROM DIFFERENT SPECIES. 151
FIGURE 4.23 PHYLOGENETIC TREE BASED ON FULL LENGTH OF HRCA PROTEIN SEQUENCES. 152
FIGURE 4.24 PSD5B SHUTTLE VECTOR. 154
FIGURE 4.25 . SYSTEMATIC PLAN OF THE FRAGMENTS CLONED INTO THE PLASMID PSD5B. 155
FIGURE 4.26 INTRINSIC SPHI SITE 31 BPS UPSTREAM TO CPN60.1 START CODON. 159
FIGURE 4.27 QUALITATIVE ASSAY WITH *M. MARINUM* STRAINS. 161
FIGURE 4.28 QUALITATIVE ASSAY WITH *M. SMEGMATIS* AND *E. COLI* STRAINS. 162
FIGURE 4.29 B-GALACTOSIDASE ASSAY IN *M. MARINUM* W.T STRAIN. 164
FIGURE 4.30 B-GALACTOSIDASE ASSAY IN *M. MARINUM* W.T AND KO STRAINS. 164
FIGURE 4.31 B-GALACTOSIDASE ASSAY IN M. MARINUM W.T AND KO STRAINS AT DIFFERENT TEMPERATURES. 167
FIGURE 4.32 B-GALACTOSIDASE ASSAY IN M. MARINUM W.T AND KO STRAINS AT DIFFERENT PH CONDITIONS. 168
FIGURE 4.33 B-GALACTOSIDASE ASSAY IN *M. SMEGMATIS*. 169
FIGURE 5.1 LIBRARY PREPARATION STEPS. 178
FIGURE 5.2 PATTERNED FLOW CELL. 178
FIGURE 5.3 CLUSTER AMPLIFICATION STEPS. 179
FIGURE 5.4 SEQUENCING BY SYNTHESIS. 180
FIGURE 5.5 MAJOR STEPS OF RNA-SEQ ANALYSIS WORKFLOW. 182
FIGURE 5.6 STEP WISE WORKFLOW OF RNA-SEQ BIOINFORMATICS DATA ANALYSIS. 188
FIGURE 5.7 PSK101 PLASMIDS (A) WITH AND (B) WITHOUT CPN10-CPN60.1. 190
FIGURE 5.8 CONSTRUCTION OF KNOCKOUT MUTANT STRAIN *M. MARINUM*. 191
FIGURE 5.9 SYMBOLIC REPRESENTATION OF M. MARINUM W.T (A), KO (B), AND COMPLEMENTED STRAIN (C). 192
FIGURE 5.10 TAPESTATION ANALYSIS REPORT SHOWING THE INTEGRITY NUMBERS OF RNA SAMPLES. 194
FIGURE 5.11 DIFFERENT STEPS INVOLVED IN RNA SEQ LIBRARY PREPARATION. 195
FIGURE 5.12 PER BASE SEQUENCE QUALITY OF THE FOUR FILES OF RNA SEQ RAW DATA. 196
FIGURE 5.13 ADAPTER CONTENT PLOT OF WT_1, R1 FILE. 197
FIGURE 5.14 BAM FILE VIEW IN ARTEMIS GENOME BROWSER. 199
FIGURE 5.15 THE RAW READ COUNTS GENERATED BY HTSEQ-COUNT. ONLY THE FIRST 21 GENES ARE SHOWN. 200
FIGURE 5.16 OUTPUT FILE OF DESEQ2 ANALYSIS BETWEEN KOVWT. ONLY THE FIRST 21 GENES ARE SHOWN. 201

FIGURE 5.17 MA PLOTS OF ALL THE THREE COMPARISONS; KOVWT (A), COMVKO (B), AND COMVWT (C). 202
FIGURE 5.18 SCATTER PLOT OF BIOLOGICAL REPLICATES WT_1 V WT_2 (A) AND WT_1 V WT_3 (B). 204
FIGURE 5.19 SCATTER PLOT BETWEEN WT_2 V WT_3 (A) AND KO_2 V KO_3 (B). 205
FIGURE 5.20 SCATTER PLOT BETWEEN KO_1 V KO_2 (A) AND KO_1 V KO_3 (B). 205
FIGURE 5.21 SCATTER PLOT BETWEEN COM_1 V COM_2 (A) AND COM_1 V COM_3 (B). 205
FIGURE 5.22 SCATTER PLOT BETWEEN COM_2 V COM_3. CREATED IN R STUDIO USING GGPLOT2 PACKAGE. 206
FIGURE 5.23 HEATMAP OF NORMALIZED COUNTS OF TRIPLICATE SAMPLES OF WT, KO, AND COM STRAINS. 207
FIGURE 5.24 SCATTER PLOT BETWEEN LOG2FOLDCHANGE VALUES OF ALL THE GENES IN KOVWT AND COMVKO COMPARISONS. CREATED IN R STUDIO USING GGPLOT2 PACKAGE. 209
FIGURE 5.25 SCATTER PLOT BETWEEN LOG2FOLDCHANGE VALUES OF ALL THE GENES IN KOVWT AND COMVWT COMPARISONS. 210
FIGURE 5.26 SCATTER PLOT BETWEEN LOG2FOLDCHANGE VALUES OF ALL THE GENES IN COMVWT AND COMVKO COMPARISONS. 210
FIGURE 5.27 HEATMAP OF NORMALIZED COUNTS OF TRIPLICATE SAMPLES OF WT, KO, AND COM STRAINS. 212
FIGURE 5.28 VENN DIAGRAMS AMONG VARIOUS COMBINATIONS OF UP AND DOWN REGULATED GENES SHOWING EXCLUSIVE AND COMMON REGULATED GENES. 214
FIGURE 5.29 SCATTER PLOT OF COMMON GENES UP IN KOVWT BUT DOWN IN COMVKO. 217
FIGURE 5.30 SCATTER PLOT OF COMMON GENES DOWN IN KOVWT BUT UP IN COMVKO. 217
FIGURE 5.31 SCATTER PLOT OF COMMON GENES DOWN IN KOVWT BUT UP IN COMVKO. 218
FIGURE 5.32 SCATTER PLOT OF COMMON GENES DOWN IN KOVWT BUT UP IN COMVKO. 218
FIGURE 5.33 VOLCANO PLOT OF KOVWT DESEQ2 ANALYSIS CREATED IN R STUDIO USING ENHANCEDVOLCANO PACKAGE. 221
FIGURE 5.34 VOLCANO PLOT OF COMVKO DESEQ2 ANALYSIS CREATED IN R STUDIO USING ENHANCEDVOLCANO PACKAGE. 222
FIGURE 5.35 VOLCANO PLOT OF COMVWT DESEQ2 ANALYSIS CREATED IN R STUDIO USING ENHANCEDVOLCANO PACKAGE. 223
FIGURE 5.36 BAR PLOT SHOWING THE TOP PATHWAYS ASSOCIATED WITH THE COMMON GENES UKVWT–DCVKO. 229
FIGURE 5.37 HIERARCHICAL CLUSTERING TREE. 229
FIGURE 5.38 BAR PLOT SHOWING THE TOP PATHWAYS. 231
FIGURE 5.39 HIERARCHICAL CLUSTERING TREE. 232
FIGURE 5.40 DIAGRAMMATIC REPRESENTATION OF CELL ENVELOPE OF MYCOBACTERIUM TUBERCULOSIS. 235
FIGURE 5.41 PEPTIDOGLYCAN BIOSYNTHESIS, DEGRADATION, AND RECYCLING PATHWAYS. 238
FIGURE 5.42 SCHEMATIC SKETCH OF THE ROLE OF FADD ENZYMES IN THE BIOSYNTHESIS OF CELL ENVELOPE PGLS AND DIMS. 239
FIGURE 5.43 SCHEMATIC DIAGRAM SHOWING THE TRANSPORT MECHANISM OF COMPLEX LIPIDS ACROSS THE CELL MEMBRANE IN *M. TUBERCULOSIS*. 239
FIGURE 5.44 BAR PLOT SHOWING THE TOP PATHWAYS ASSOCIATED WITH THE COMMON GENES DKVWT – UCVKO. 241
FIGURE 5.45 THE HIERARCHICAL CLUSTERING TREE SHOWING THE CORRELATION OF DIFFERENT PATHWAYS SHOWN IN FIGURE 5.41. 241
FIGURE 5.46 RELATIVE EXPRESSION OF CHAPERONIN GENES IN WT SAMPLES OF M. MARINUM 1218R. 244
FIGURE 5.47 RELATIVE EXPRESSION OF CHAPERONIN GENES IN WT SAMPLES OF M. TUBERCULOSIS H37RV. 244
FIGURE 6.1 MAP OF CLONING AND EXPRESSION VECTOR PTRC99A. 253
FIGURE 6.2 GENETIC ORGANIZATION OF GROESL REGION AND SURROUNDING CDS OF E. COLI MG1655 AND MGM100 STRAINS. 254
FIGURE 6.3 SCHEMATIC DIAGRAM SHOWING THE PLASMID SHUFFLING STRATEGY IN *E. COLI* AI90 STRAIN. 255
FIGURE 6.4 SCHEMATIC DIAGRAM OF DIFFERENT STEPS OF CLONING OF *GROES*, MM*CPN60.1*, AND MM*CPN60.2* INTO PTRC99A. 256

FIGURE 6.5 SCHEMATIC DIAGRAM OF DIFFERENT STEPS OF CLONING OF THE MMCPN10, MMCPN60.1 AND MMCPN60.2 INTO PTRC99A. 257
FIGURE 6.6 COMPLEMENTATION ASSAY IN MGM100 STRAIN. 261
FIGURE 6.7 COMPLEMENTATION ASSAY USING MGM100 STRAIN TO FIND THE BEST WORKING CONCENTRATION OF IPTG. 262
FIGURE 6.8 SDS-PAGE GEL. 264
FIGURE 6.9 COMPLEMENTATION ASSAY IN E. COLI AI90 STRAIN. 266
FIGURE 6.10 GRAPHS SHOWING THE DISTRIBUTION OF CODON USAGE FREQUENCY ALONG THE LENGTH OF THE GENE SEQUENCES. 268
FIGURE 6.11 PERCENTAGE DISTRIBUTION OF CODONS OF MM*CPN60.1* AND MM*CPN10*. 268
FIGURE 6.12 PERCENTAGE DISTRIBUTION OF CODONS OF CODON OPTIMIZED MMCPN60.1 AND MMCPN10. 269
FIGURE 6.13 PERCENTAGE DISTRIBUTION OF CODONS OF OPTIMIZED GENE SEQUENCES. 269
FIGURE 6.14 RBS DISTANCE OPTIMIZATION. 270
FIGURE 6.15 COMPLEMENTATION ASSAY IN TAB21 AT 30°C. 271
FIGURE 6.16 SDS-PAGE GEL. 271
FIGURE 6.17 COMPLEMENTATION ASSAY IN MGM100 STRAIN. 273
FIGURE 6.18 SCHEMATIC DIAGRAM OF DIFFERENT STEPS OF CLONING FOR CONSTRUCTION OF PTRC-MMCPN10-GROEL PLASMID. 274
FIGURE 7.1 COMPLEMENTATION ANALYSIS OF SINGLE AND DOUBLE PHOSPHOMIMETIC MUTANTS OF MTBCPN60.1. 281
FIGURE 7.2 COMPLEMENTATION ANALYSIS OF DOUBLE PHOSPHOMIMETIC PLUS ATPASE MUTANTS OF MTBCPN60.1. 282
FIGURE 7.3 1-MULTIPLE SEQUENCE ALIGNMENT OF CPN60.2 AND CPN60.1 OF TEN SPECIES OF MYCOBACTERIA. 285
FIGURE 7.4 2-MULTIPLE SEQUENCE ALIGNMENT OF CPN60.2 AND CPN60.1 OF TEN SPECIES OF MYCOBACTERIA. 286
FIGURE 7.5 3-MULTIPLE SEQUENCE ALIGNMENT OF CPN60.2 AND CPN60.1 OF TEN SPECIES OF MYCOBACTERIA. 287
FIGURE 7.6 SERINE/THREONINE PROTEIN KINASES OF M. TUBERCULOSIS H37RV AND M. MARINUM 1218R GENOME. 290
FIGURE 7.7 VISUALIZATION OF PKNF OF H37RV AND THE CORRESPONDING GENE OF 1218R IN ARTEMIS COMPARISON TOOL. 292
FIGURE 7.8 COMPLEMENTATION ASSAY OF SINGLE AND DOUBLE PHOSPHOMIMETIC MUTANTS OF MMCPN60.1-CO. 293
FIGURE 7.9 OVEREXPRESSION OF DOUBLE RBS OPTIMIZED PHOSPHOMIMETIC MUTANTS OF M. MARINUM CPN60.1-CO. 294
FIGURE 7.10 COMPLEMENTATION ASSAY OF VARIOUS DOUBLE PHOSPHOMIMETIC MUTANTS OF MMCPN60.1-CO. 295
FIGURE 7.11 COMPLEMENTATION ANALYSIS OF WT-CO AND DPM OF MMCPN60.1 IN COMBINATION WITH E. COLI GROES. 297
FIGURE 7.12 WESTERN BLOT OF MMCPN60.1-CO-DPM EXPRESSED FROM PLASMIDS IN MGM100 STRAIN. 299
FIGURE 7.13 COMPLEMENTATION ASSAY OF ATPASE MUTANTS IN MGM100. 302
FIGURE 7.14 SEQUENCE ALIGNMENT OF *M. TUBERCULOSIS* CPN60.1, *M. BOVIS* CPN60.1 AND *M. MARINUM* CPN60.1 PROTEINS. 305
FIGURE 7.15 SCHEMATIC PLAN FOR SWAPPING OF HINDIII FRAGMENT BETWEEN MTB*CPN60.1*-DPM AND MM*CPN60.1*-DPM. 306
FIGURE 7.16 COMPLEMENTATION ASSAY IN MGM100. 309
FIGURE 7.17 WESTERN BLOT OF MMCPN60.1 PURIFIED USING IMMUNOPRECIPITATION AND NI-NTA MAGNETIC BEADS. 311
FIGURE 7.18 RESULTS OF MASS SPECTROMETRY ANALYSIS OF MMCPN60.1. 311

List of Tables

TABLE 1.1 DIFFERENT CLASSES OF MOLECULAR CHAPERONES AND THEIR EXAMPLES IN BACTERIA, ARCHAEA, AND EUKARYOTIC CELLS. 21
TABLE 1.2 DIFFERENT CLASSES OF MOLECULAR CHAPERONES AND THEIR EXAMPLES IN *E. COLI*, *M. MARINUM*, AND *M. TUBERCULOSIS.* 22
TABLE 2.1 LIST OF BACTERIAL STRAINS USED IN THIS STUDY. 44
TABLE 2.2 TABLE OF PLASMIDS USED IN THIS STUDY. 49
TABLE 2.3 PRIMERS USED FOR AMPLIFICATION OF UPSTREAM REGIONS OF *M. MARINUM* 1218R CPN10, CPN60.1 AND CPN60.2 FOR CLONING INTO PSD5B. 50
TABLE 2.4 OLIGOS FOR INSERTION OF MISSING 31 BPS IN P5B_60.1_S AND P5B_60.1_L PLASMID CONSTRUCTS (FIGURE 4.23). 51
TABLE 2.5 PRIMERS USED FOR AMPLIFICATION OF CPN10 GENE FROM *M. MARINUM* GENOMIC DNA TO CLONE INTO PTRC99A BETWEEN NCOI AND BGLII SITES. (FIGURE 6.5). 51
TABLE 2.6 PRIMERS USED FOR REMOVAL OF INTRINSIC NCOI SITE FROM CPN60.1 GENE OF *M. MARINUM* IN PTRC-GROES-MMCPN60.1. 51
TABLE 2.7 PRIMERS USED FOR SITE DIRECTED MUTAGENESIS TO CHANGE THE DISTANCE OF RBS IN PTRC- MMCPN10-MMCPN60.1 CO (FIGURE 6.15). 52
TABLE 2.8 PRIMERS USED FOR CONSTRUCTION OF PHOSPHOMIMETIC MUTATIONS. 54
TABLE 2.9 PRIMERS USED FOR CONSTRUCTION OF ATPASE MUTATIONS. 54
TABLE 2.10 PRIMERS USED FOR SEQUENCING AND SCREENING. 56
TABLE 2.11 PRIMERS USED SDM TO INSERT BGLII SITE (CHAPTER 6, FIGURE 6.17). 56
TABLE 2.12 STOCK AND WORKING CONCENTRATIONS OF SUPPLEMENTS USED IN THIS STUDY. 58
TABLE 2.13 REAGENTS FOR INOUE TRANSFORMATION BUFFER. 64
TABLE 2.14 REAGENTS USED FOR THE PREPARATION OF LACZ BUFFER. 68
TABLE 2.15 RESTRICTION DIGESTION PARAMETERS. 70
TABLE 2.16 PCR REACTION SET UP FOR AMPLIFICATION OF FRAGMENTS FROM GENOMIC DNA. 71
TABLE 2.17 PARAMETERS USED FOR PCR AMPLIFICATION. 72
TABLE 2.18 REAGENTS USED FOR THE PHOSPHORYLATION OF OLIGOS. 72
TABLE 2.19 PCR REACTION SET UP FOR SDM. 73
TABLE 2.20 PCR PARAMETERS FOR SITE-DIRECTED MUTAGENESIS. 73
TABLE 2.21 REAGENTS USED FOR THE PREPARATION OF SDS 5X LOADING BUFFER. 74
TABLE 2.22 REAGENTS USED FOR THE PREPARATION OF SDS-PAGE GEL. 75
TABLE 3.1 TABLE SHOWING THE CORE AND ACCESSORY GENES OBTAINED BY PANGENOME ANALYSES OF 1218R, H37RV, AND MC2155. 98
TABLE 3.2 NAMES OF THE SPECIES USED FOR PHYLOGENETIC AND SEQUENCE ANALYSIS. 115
TABLE 5.1 TABLE SHOWING THE ABSORBANCE AND RIN NUMBER. 193
TABLE 5.2 NUMBER OF READS AND OVERALL ALIGNMENT RATE. 198
TABLE 5.3 THE LOG2FOLDCHANGE AND PADJ VALUES OF A FEW OUTLIERS IN THE SCATTER PLOTS OF KO SAMPLES. 204
TABLE 5.4 THE LOG2FOLDCHANGE AND PADJ VALUES OF A FEW OUTLIERS IN THE SCATTER PLOTS 5.22. 208
TABLE 5.5 TWO SETS OF DIFFERENTIALLY EXPRESSED GENES SORTED BY LOG2FOLDCHANGE AND PADJ VALUES OF THE THREE COMPARISONS OF DESEQ2 ANALYSIS. 211
TABLE 5.6 LIST OF COMMON AND EXCLUSIVE GENES IN KOVWT AND COMVKO COMPARISONS. 215
TABLE 5.7 SIX GENES COMMON DOWNREGULATING IN KOVWT AND UPREGULATING IN COMVWT COMPARISONS. 215
TABLE 5.8 TABLE SHOWING THE NAMES, LOG2FOLDCHANGE, AND PADJ VALUES OF THE OUTLIERS IN FIGURE 5.29 AND 5.30. 219
TABLE 5.9 DETAIL OF PROTEINS ASSOCIATED WITH MOLECULAR CHAPERONES, STRESS RESPONSE AND PROTEIN FOLDING PATHWAYS. 227
TABLE 5.10 GO PATHWAY ANALYSIS PERFORMED OF COMMON GENES SIGNIFICANTLY UKVWT _DCVKO. 228
TABLE 5.11 DETAIL OF PROTEINS ASSOCIATED WITH CELL WALL, CELL DIVISION, AND PEPTIDOGLYCAN PATHWAYS. 233
TABLE 5.12 DETAIL OF PROTEINS ASSOCIATED WITH POLYKETIDE SYNTHASE, PHOSPHOPANTETHEINE-BINDING, AND RELATED PATHWAYS. 234
TABLE 5.13 TOP 10 PATHWAYS ACCORDING TO THE FOLD ENRICHMENT. 242

TABLE 6.1 SUMMARY OF THE COMPLEMENTATION ASSAY IN E. COLI MGM100. 263
TABLE 7.1 SERINE/THREONINE PROTEIN KINASES OF M. TUBERCULOSIS H37RV AND M. MARINUM 1218R. 291
TABLE 7.2 ATPASE MUTATIONS CREATED IN MMCPN60.1-CO (W.T, DPM, DPM-8RBS). 300

Abbreviations

CO	Codon optimized (sequence optimized)
Cm^R	Chloramphenicol resistance
Cpn	chaperonin
DPM	Double phosphomimetic mutant (S75D, S393D)
DMSO	Dimethyl sulfoxide
DMF	Dimethylformamide
DW	Distilled water
ddH_2O	Double distilled water
EDTA	Ethylenediamine tetraacetic acid
GTE	Glucose/Tris/EDTA
GTF	Gene transfer format
Hyg^R	Hygromycin resistance
IPTG	Isopropyl ß-D-1-thiogalactopyranoside
IgG	Immunoglobulin G
Kan^R	Kanamycin resistance
LB	Luria-Bertani
$MgCl_2$	Magnesium chloride
$CaCl_2$	Calcium chloride
$MgSO_4$	Magnesium sulphate
Mtb	*Mycobacterium tuberculosis*
Mm	*Mycobacterium marinum*
Ms	*Mycobacterium smegmatis*
MW	Molecular weight
OD	Optical density
ONPG	ortho-Nitrophenyl-β-galactoside
PAGE	Polyacrylamide gel electrophoresis

PBS	Phosphate buffered saline
PIPES	piperazine-N, N′-bis 2-ethanesulfonic acid
PROKKA	Rapid Prokaryotic Genome Annotation
psi	pound per square inch (unit of pressure)
PVDF	Polyvinylidene fluoride or polyvinylidene difluoride
RBS	Ribosome binding site
RIN	RNA integrity number
Roary	High speed stand-alone pan genome pipeline
SDM	Site directed mutagenesis
SDS	Sodium dodecyl sulphate
SOB	Super Optimal broth
SOC	Super Optimal broth with Catabolite repression
Str^R	Streptomycin resistance
STPKs	Serine/threonine protein kinases
TBST	Tris Buffered Saline with Tween
Tet^R	Tetracycline resistance
TAE	Tris-acetate-EDTA
6RBS	6 bps between RBS and start codon
8RBS	8 bps between RBS and start codon
10RBS	10 bps between RBS and start codon
X-Gal	5-bromo-4-chloro-3-indolyl-β-D-galactoside

1 Introduction

1.1 Tuberculosis

Tuberculosis (TB) is an ancient lethal disease caused by a BSL3 level pathogen *Mycobacterium tuberculosis*. Historically, tuberculosis is thought to be as ancient as 3000 - 1500 BC (Barberis et al., 2017). In the 18th century, tuberculosis had become a seasonal epidemic in Europe and North America (Zürcher et al., 2016) which killed nearly 900 people per one hundred thousand population per year (Daniel, 2006). In the 19th century, tuberculosis killed nearly one fourth of the adult population in Europe. Between the years 1851 to 1910, approximately 4 million people died of tuberculosis in England and Wales, half of which were youngsters aged 20 – 24 (Zürcher et al., 2016). By the end of 19th century nearly 90% of the metropolitan inhabitants of Europe and United States were infected with *M. tuberculosis* (John, 2021).

Discovery of the aetiological agent of human tuberculosis (TB) was first announced by Robert Koch on 24 March 1882 (Jeremiah et al., 2022). The first successful vaccine made from the attenuated bovine strain of tuberculosis (*M. bovis*), called BCG, was used on human in 1921. By 1956 nearly 600,000 people had been vaccinated (Andersen & Doherty, 2005). After the first isolation of streptomycin from *Streptomyces griseus* in 1944, it was found to be the first effective antibiotic against *Mycobacterium tuberculosis* (Kingston, 2004). In 1952 the first oral antibiotic, isoniazid, was developed. Later, the emergence of rifampicin in 1970s revolutionized the treatment of tuberculosis (Vilchèze & Jacobs, 2019).

The hopes for the elimination of tuberculosis disease were ruined in 1980s due to the emergence of drug-resistant strains. The tuberculosis cases started to rise again because of the rise of multi-drug resistant (MDR-TB). MDR-TB is resistant to the two most effective antibiotics for tuberculosis i.e., isoniazid and rifampicin. In 1990 approximately 2.6 – 2.9 million died from tuberculosis worldwide. The magnitude of problem increased with increase in tuberculosis cases and deaths in HIV patients (Styblo, 1990). Tuberculosis is reported to be the reason of 30% of AIDS related deaths. World health organization (WHO) declared TB as a global health emergency in 1993 (Dooley, 1992; Grange & Zumla, 2002). Although because of the availability of first- and second-line antibiotics, tuberculosis is a curable disease. The emergence of extremely drug resistant (XDR-TB) strain that is resistant to isoniazid, rifampicin,

fluoroquinolone, and at least one of the injectable drugs for tuberculosis has exaggerated the problem (Balabanova et al., 2016; Khawbung et al., 2021).

In the past two decades the global tuberculosis cases and deaths have been declining slowly, but the rate is not fast enough to meet the goal of end TB strategy of WHO (Figure 1.1). In 2020, 9.9 million active tuberculosis cases very reported worldwide. A drop in the notification of new cases of TB in 2020 was due to COVID-19 pandemic which has reversed the years of global efforts in decreasing the TB related deaths. Approximately 1.3 million people died from tuberculosis alone in 2020 with an additional 214,000 HIV-positive TB deaths reported worldwide. Currently, TB is 13th top reason of death globally and a second top contagious disease after COVID-19 (World Health Organization (WHO), 2021).

One reason of increase in number of deaths from TB is increase in multi-drug resistance (MDR-TB) and rifampicin resistance (RR-TB) tuberculosis cases (Figure 1.2). The recovery rates for MDR-TB infections are only 50% – 60% (Chung-Delgado et al., 2015; Koul et al., 2011). Figure 1.2 shows the percentage of treatment results of new and relapsed tuberculosis cases and MDR/RR-TB cases from 2012 – 2019 worldwide. The death rate of MDR/RR-TB patients was recorded to be much higher than the normal TB cases. The treatment success of non-resistant TB cases was nearly 20% higher than the drug resistant cases (Figure 1.2).

The key obstacle in the eradication of TB is not only drug resistance but also the development of latent tuberculosis infection which is discussed in the next section. The development of non-antibiotic treatments which target the survival mechanisms of *M. tuberculosis* (Mtb) inside the host and help the clearing of infection is indispensable. To find the novel molecular targets for treatment of tuberculosis, a basic understanding of tuberculosis pathogenesis is required. The development mechanism of tuberculosis is briefly described in the following section.

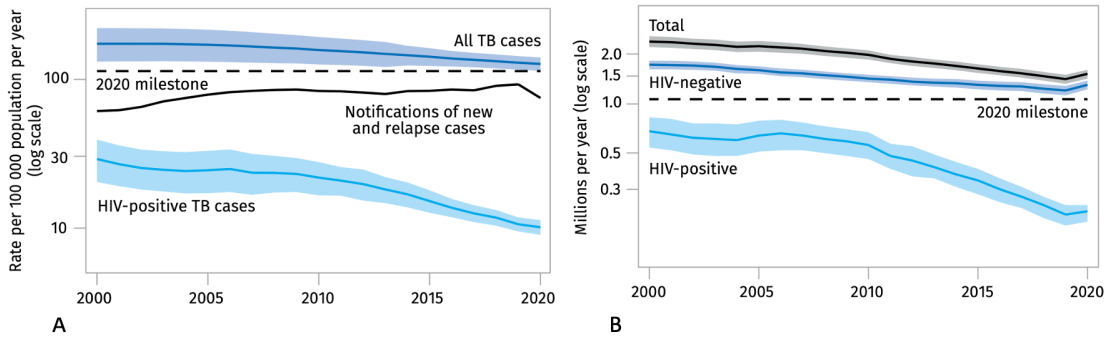


Figure 1.1 The estimated number of reported tuberculosis cases (A) and deaths (B) worldwide from the year 2000 – 2020.
The dashed line in (B) shows the 2020 target set by the End TB strategy of WHO. Shaded areas indicate uncertainty intervals. Images are taken from global tuberculosis report – 2021, WHO.

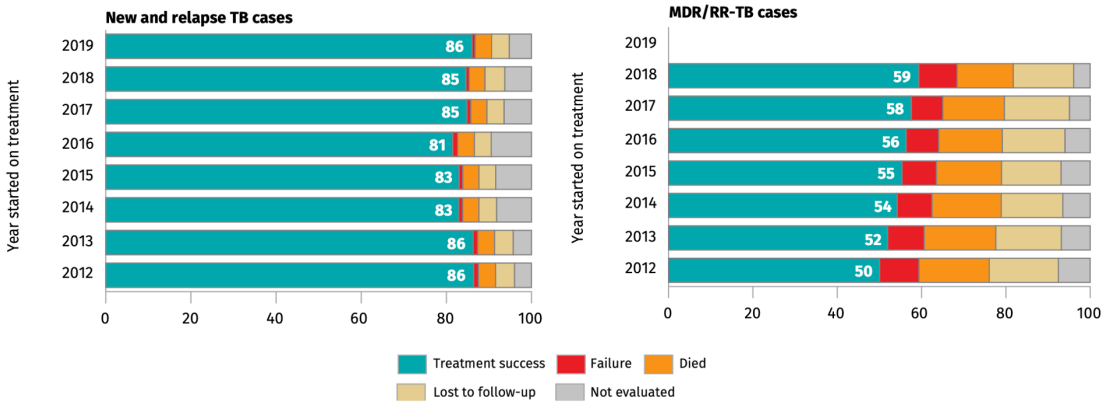


Figure 1.2 Treatment results of new and relapse tuberculosis cases and MDR-TB / RR-TB cases worldwide from the year 2012 – 2019.
Images are taken from global tuberculosis report – 2021, WHO.

1.1.1 *M. tuberculosis* infection

M. tuberculosis spreads by the coughing and sneezing of the active tuberculosis patients through air. Upon inhalation by a healthy person, the bacteria reach the alveoli of the lungs where they are engulfed by alveolar macrophages which are the fundamental components of innate immune response of the host (Myllymäki et al., 2016). The phagocytosed *M. tuberculosis* use various strategies to escape the immune response of the host such as expression of virulence factors, blocking of fusion of phagosomes with lysosomes, defence against reactive oxygen species (ROS), prevention of phagosome maturation, and inhibition of apoptosis (Shim et al., 2020).

The infection of alveolar macrophages with bacilli lead to the induction of a chain of proinflammatory responses resulting in the recruitment of leukocytes such as dendritic cells, natural killer cells, lymphocytes, neutrophils, and additional macrophages to the pulmonary infection site (Myllymäki et al., 2016; Russell et al., 2009; Silva Miranda et al., 2012). Accumulation of leukocytes around the infected macrophages because of signalling events of the immune system develop a matrix of early-stage granuloma which is the hallmark feature of tuberculosis (Figure 1.3).

In the early stages of granuloma formation, the infected macrophages are present in the centre of a circle made by the foamy macrophages and monocytes which are surrounded by an external circle of lymphocytes (Russell et al., 2009). Granulomas are formed by the host immune cells to create a microenvironment to control the infection. The bacteria can survive inside the granulomas leading to the latent tuberculosis infection (LTBI) (Flynn et al., 2011; Shim et al., 2020). As the granulomas start to mature an extensive fibrous casing is formed around the foamy macrophages. In the centre of the granuloma due to necrosis of the infected macrophages a soft core of necrotic zone containing cell debris is formed that acquire cheese-like consistency therefore called caseous granuloma (Figure 1.3) (Russell et al., 2009; Sinigaglia et al., 2020). This infection can remain localized where bacteria can survive and persist inside the granuloma for the lifetime in immunocompetent hosts. However, 10% of the latent infections subsequently reactivate. The centre of the granuloma liquefies and bacilli multiply rapidly leading to the eruption of the granuloma (Rao et al., 2019).

Different types of necrotic granulomas have been observed for example., fibrotic, calcified, and suppurative. The fibrotic granulomas are encapsulated by a fibrotic scaffold, calcified granulomas have mineralized necrotic core whereas in the suppurative granulomas neutrophilic debris and neutrophils are found in the centre of the necrotic core. The morphology of primary granulomas is also distinct from the mature granulomas (Lenaerts et al., 2015; Mattila et al., 2013). As a result of *M. tuberculosis* infection non-necrotic granulomas are also produced where bacilli inhabit inside the infected immune cells. This type of granulomas is called cellular granulomas. Although significant heterogeneity has been observed in the different types of granulomas, a general pattern is found to be uniform among all types of granulomas i.e., macrophages are generally present in the centre and first rim of the granulomas whereas lymphocytes are present at the periphery forming the outer rim of the granulomatous structure (Kauffman et al., 2018; Mattila et al., 2013; Ramakrishnan, 2012).

Granulomas are found both in latent and active tuberculosis infections therefore, the formation of granuloma does not guarantee the control of bacterial infection. In active tuberculosis, granulomas rupture and spill out thousands of bacilli which spread around and lead to formation of more granulomas (Delogu et al., 2013; Russell et al., 2009). The knowledge of latent tuberculosis infection is limited and the full knowledge of bacterial and/or host factors leading to granuloma formation is not understood well. Only a few localized granulomas are found in lungs in case of latent infection which limit the spread and multiplication of *M. tuberculosis* (Flynn et al., 2011).

About 5% of tuberculosis infections are cleared up by the host and does not lead to latent or active infection. Nearly 90 – 95% infections result in latent tuberculosis infection with 5% risk of developing active TB in first 2 years of the infection and 10% lifetime risk of progression to active TB. Only 5% tuberculosis infections directly develop active disease. Although effective antibiotic treatment is available for active TB, but the success of the treatment depends on many factors such as the resistance level of the infection strain and immunocompetency level of the patient (Figure 1.4) (Jeremiah et al., 2022; Koul et al., 2011; World Health Organization (WHO), 2021).

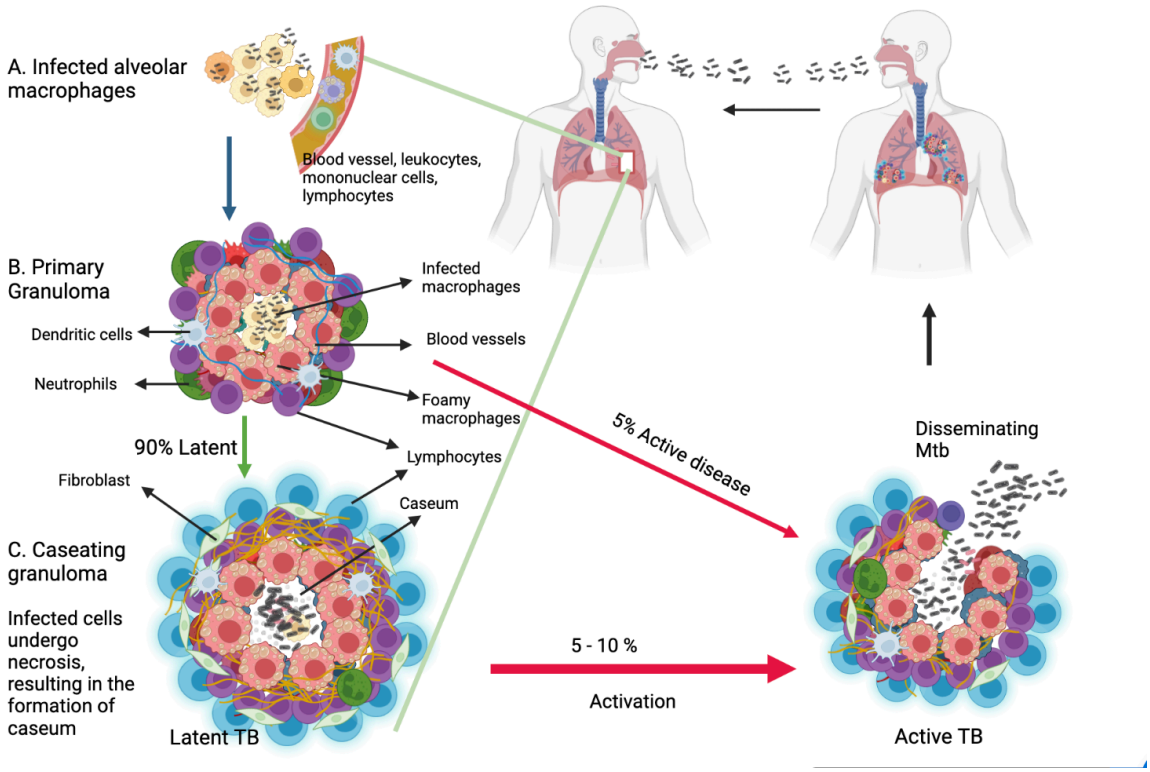


Figure 1.3 Granuloma formation and progression in the lungs of human following inhalation of *M. tuberculosis* contaminated aerosols.
Figure adapted from (Myllymäki et al., 2016; Russell et al., 2009).

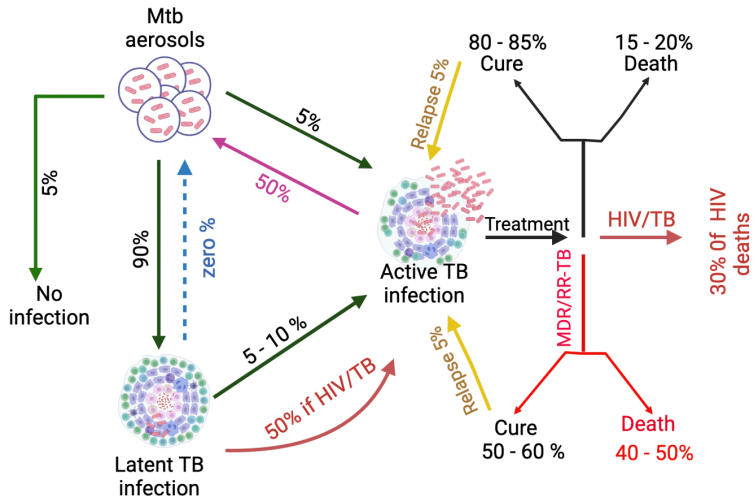


Figure 1.4 Different types of *M. tuberculosis* infection and percentage of the risk of each type. *M. tuberculosis* spreads through aerosols. The person inhaling the Mtb droplets have 90% chance of developing latent TB infection with 5% chance of progression to active TB in first 2 years of the infection and 10% lifetime risk to develop active TB. The first- and second-line antibiotic treatment is available which can cure TB in 80-85% of cases but if the infection is caused by MDR/RR-TB strain then chances of death are increased up to 50%. A person with active TB infection can spread the disease by coughing and sneezing but TB is not contagious in latent infection condition (Koul et al., 2011; World Health Organization (WHO), 2021).

Our understanding of the development of granulomas is very limited. For example, what type of granulomas lead to a lifetime latent infection and what factors cause the reactivation of latent infection to active TB. We may call a granuloma functioning if it could eradicate the infecting bacilli or control the bacteria for lifetime. There could be multiple factors and responses from the *M. tuberculosis* and host that establish a functioning granuloma.

According to the latest estimates of WHO one fourth of the world population is infected with latent TB infection. Drug resistant *M. tuberculosis* strains and latent TB are key obstacles in elimination of TB. The development of non-antibiotic treatments of tuberculosis is required which could target the molecular weapons and defence mechanism of *M. tuberculosis*. A deeper understanding of host-Mtb interactions, granuloma formation and discovery of novel molecular targets could lead to achieving the goal of eradication of tuberculosis.

1.1.2 Determinants of granuloma formation - survival and proliferation of *M. tuberculosis* in host

Studies have shown that *M. tuberculosis* is able to replicate and propagate to limited extent within the granulomas. The inside environment of granulomas provides various challenging stresses to the bacilli such as low pH, nutrient starvation, hypoxia, iron limitation, and lethal levels of reactive oxygen and nitrogen species. *M. tuberculosis* adapt several mechanisms to survive in this challenging environment. For example, to survive in hypoxia *M. tuberculosis* utilize DosR regulon which is a set of nearly 48 coregulated genes induced in the hypoxic conditions.

Dual staining of the lung sections of *M. tuberculosis* infected lungs of mice and guinea pig suggested that the intracellular bacilli had a distinct pattern of cell wall lipids compared to those grown in 7H9 media. In a single microenvironment multiple populations of *M. tuberculosis* with altered cell wall components were observed (Ehlers & Schaible, 2013; Fischer et al., 2001; Ryan et al., 2010). Lenaerts et al., 2007 performed the acid-fast staining and tissue histology of lungs of *M. tuberculosis* infected guinea pigs after treatment with isoniazid, rifampicin, and pyrazinamide. A large number of extracellular bacilli forming clusters were found at the rim region around the necrotic

core. All the acid-fast bacilli were present outside the immune cells forming structures similar to biofilms.

Chakraborty et al., 2021 studied the structural components of biofilms of *Mycobacterium avium*, *Mycobacterium abscessus*, and *Mycobacterium fortuitum*. Cellulose was found to be a crucial structural component of all the three mycobacterial biofilms. *M. tuberculosis* biofilm like structures enclosed in extracellular cellulose matrix were also observed in the lungs of mouse and macaque monkey's lungs infected with *M. tuberculosis*. Similar results were found in the lungs sections of humans died from *M. tuberculosis* (Chakraborty et al., 2021). Several clusters of *M. tuberculosis* bacilli enclosed in the cellulose matrix were observed in the lungs of mouse, macaque monkey, and humans infected with Mtb. The chemotherapy of *M. tuberculosis* infected mice showed that disruption of cellulose by nebulization of cellulase increased the efficacy of isoniazid and rifampicin (Chakraborty et al., 2021).

Biofilms of *M. tuberculosis* are yet to be studied. However, in *Mycobacterium smegmatis* biofilm formation, a non-essential chaperonin Cpn60.1 was observed to be essential for mycolic acid biosynthesis and biofilm maturation. It was found that *M. smegmatis* under regulation of Cpn60.1 built up unusual short-chain fatty acids thus altering the mycolate profile which is required for biofilm formation. This activity is not required for the normal planktonic growth of *M. smegmatis* (Ojha et al., 2005; Röse et al., 2004; Teng & Dick, 2003; Zambrano & Kolter, 2005).

Another study with WT and Δ*cpn60.1* and complemented strains of *Mycobacterium bovis* demonstrated that the WT and complemented strains developed a fully mature biofilm in 6% Sauton's media whereas, the biofilm of KO strain was impaired and defective. This could partly be because Δ*cpn60.1* could not produce PDIM (a lipid, phthiocerol dimycocerosates) which is a component of mycobacteria cell wall. And very low amount of PGL (phenolic glycolipids) were observed to be produced in the KO (Δ*cpn60.1*) strain. The extent of effect of absence of PDIM alone or PDIM and PGL together in the presence of Cpn60.1 on biofilm formation was only negligible indicating that the poor biofilm formation in Δ*cpn60.1 M. bovis* strain was not completely associated with the PDIM/PGL production. This shows the importance of chaperonin Cpn60.1 in cell wall biogenesis and in biofilm formation as well (Zeng et al., 2019).

M. tuberculosis is the world's most successful pathogen which exploits and subverts the host's immune response for its own survival. Scientists have been thinking that the granuloma formation is driven by the immune response of the host for protection against the invading pathogen. However, recent studies have shown that the proliferation of *M. tuberculosis* is connected with the formation of granulomas and that the determinants of granuloma formation come from the bacilli (Davis & Ramakrishnan, 2009; Volkman et al., 2004).

Studies have shown the importance of cell wall lipids of *M. tuberculosis* in granuloma formation. Several specific types of lipids were found to be major moderators for a robust granulomatous response. Trehalose 6, 6'-dimycolate (TDM – a component of mycolic acid cell wall) alone was sufficient to trigger the granuloma formation; intraperitoneal injections of TDM in sensitized mice (low-dose of *M. tuberculosis*) produced robust caseating granulomas. TDM was also observed to be crucial for granuloma formation as the removal of Trehalose 6, 6'-dimycolate, lipids and other mycolic acid components from the cell surface of *M. tuberculosis* using lipid extraction technique abolished the granuloma formation in infected mice (Chang et al., 1986; Hunter et al., 2006; Sabin, 1932; C. L. Silva et al., 1985).

In another study the chaperonin Cpn60.1 was found to be required for granuloma formation. The study was done using WT, Δ*cpn60.1* (KO), and complemented strains of *M. tuberculosis*. 15 weeks post infection the mice infected with WT strain developed large areas (covering nearly 50% of the entire lungs) of granulomatous inflammation. Whereas the mice infected with KO (Δ*cpn60.1*) strain showed only insignificant increase in the inflammatory cells in the alveolar region. Complemented strain showed granulomatous inflammation nearly equal to the level of WT strain. No difference in the lipid profiles of WT and KO strains grown in liquid or solid media (*in vitro*) was observed (Hu et al., 2008). Similar results were found with guinea pig infected with WT, KO (Δ*cpn60.1*) and complemented strains. 12 weeks post infection, the lungs of animals infected with WT and complemented strain were found to have large number of accumulated inflammatory cells including macrophages, neutrophils, and lymphocytes. However, the lungs of the animals infected with KO strain were mostly normal with only negligible number of inflammatory cells present (Hu et al., 2008). The cytokines expression of animals infected with KO strain was considerably reduced

compared to the infection with WT and complemented strains. Both experiments showed that the Δ*cpn60.1* strain of *M. tuberculosis* was less inflammatory and lacked the ability to induce granuloma formation.

Although granuloma formation seems to be an effective host response against *M. tuberculosis* which keeps the infection in latent state in 90% of the cases but rest of the individuals which develop active tuberculosis and the 5 – 10 % risk of progression of latent TB to active TB raise many questions about the granulomas to be a successful mechanism of host defence. The risk of latent TB progression to active TB even increases to 50% in case of immunocompromised individuals. The study of molecular determinants of *M. tuberculosis* involved in driving the granuloma formation and the survival mechanisms of the bacilli inside the granuloma could provide a better understanding of tuberculosis pathogenesis and novel biomarkers for non-antibiotic treatment of active and latent tuberculosis.

As discussed above, in many studies a chaperonin Cpn60.1 of mycobacteria was found to play an important role in the cell wall biogenesis and biofilm maturation of *M. smegmatis* and *M. bovis.* Also, the Cpn60.1 of *M. tuberculosis* was found to be crucial for granulomatous inflammation. The summary of different studies demonstrating the role of Cpn60.1 and cell wall mycolic acid in granuloma formation is given in Figure 1.5. The chaperonins of mycobacteria seem to play an important role in many vital and crucial mechanisms of the cell. In previous studies, *M. smegmatis* and *M. bovis* have been largely used as model organism to understand the pathogenesis of Mtb. We used *M. marinum* as a model organism to study the mechanism and functions of chaperonins of *M. tuberculosis*. In our laboratory zebrafish – *M. marinum* model is being used to study the effect of different mutants on pathogenesis and granuloma formation. The detail of zebrafish – *M. marinum* model is discussed in next section and the detailed introduction of chaperonins of mycobacteria is given in section 1.4.

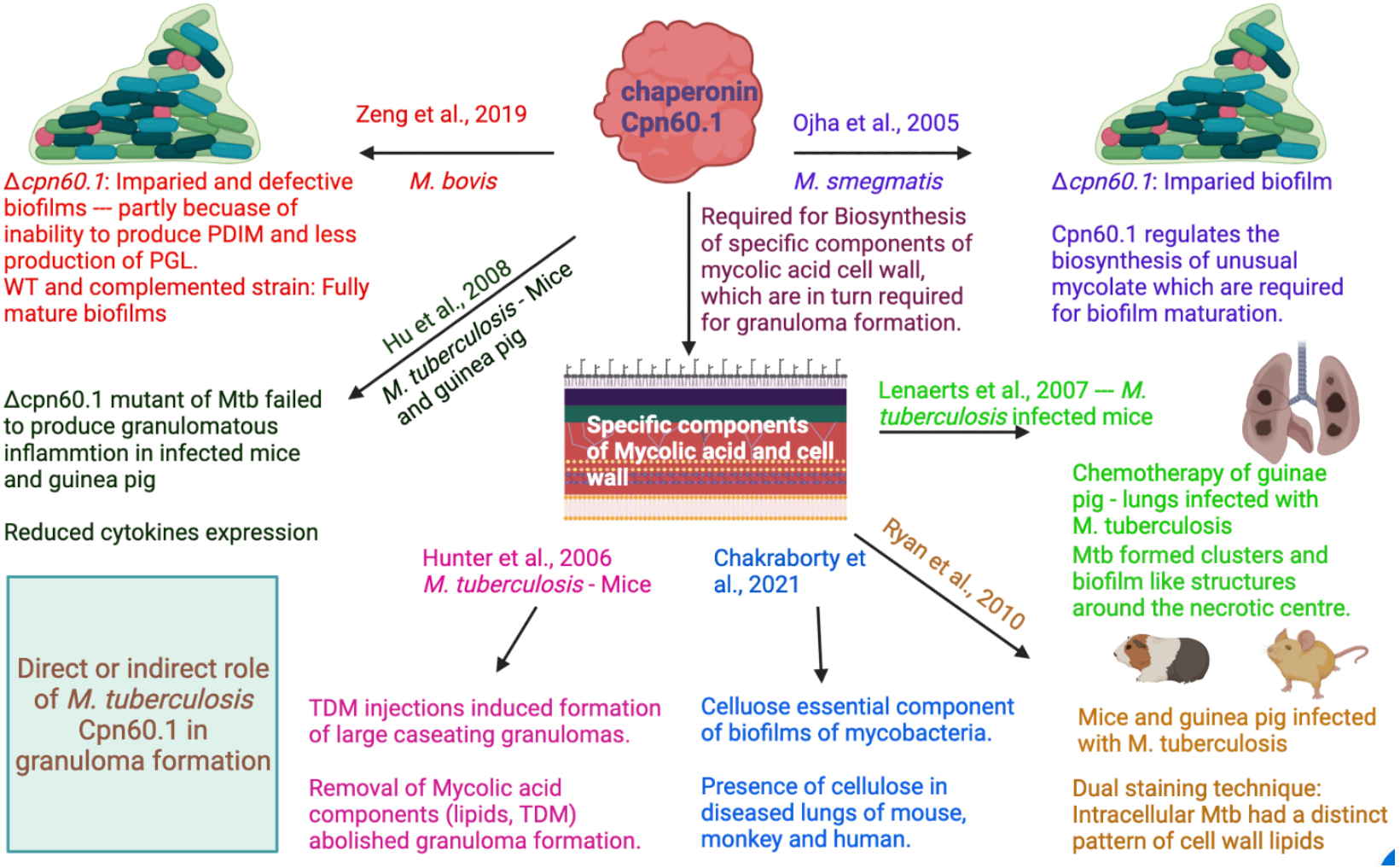


Figure 1.5 A summary of different studies demonstrating the role of some bacterial determinants in granuloma formation, the survival and proliferation of *M. tuberculosis* in the stressed environment of granulomas.

1.2 Zebrafish - *Mycobacterium marinum* model

1.2.1 Zebrafish (Danio rerio) – A better host to study Tuberculosis

About a century ago, zebrafish embryos attracted the attention of developmental biologists because of their optical transparency and embryological manipulability. In 1930s, zebrafish embryos and larvae were being used as a model for classical developmental and embryological studies (Lieschke & Currie, 2007). Zebrafish embryogenesis have been extensively used for understanding mechanisms of development since mid of the twentieth century because the embryos are external, transparent, inexpensive, and readily maintainable. The chemical and genetic tractability of zebrafish was discovered in 1980s which breathed a new life in the study of host-pathogen interaction. Many approaches for the genetic manipulation of zebrafish such as cloning, transgenesis, and mutagenesis were developed (Lawson & Wolfe, 2011). Several zebrafish disease models were raised using both forward and reverse genetics approaches. The possibility of genome editing strengthened the use of zebrafish as an invertebrate-style forward genetics model for the study of several human diseases. Research using zebrafish exploded in the 1990s when many successful saturation mutagenesis and genetic screening methods were developed, and thousands of early development zebrafish mutants were identified (Amsterdam et al., 1999; Driever et al., 1996). Nowadays zebrafish is an extensively used vertebrate animal model to study several human diseases from cancer, nervous and cardiac disorders to inflammatory infections and immunology studies.

The full sequence, assembly and complete annotation of zebrafish genome show approximately 70% direct orthology with human reference genome, including disease-associated genes (Howe et al., 2013). Zebrafish reference genome sequence has led to practicable genome engineering and efficient generation of stable mutants. A single gravid zebrafish female can produce around 150 – 400 eggs per spawning. Embryos are more susceptible to genetic engineering providing more feasible genetic screening compared to adult zebrafish (Blackburn et al., 2013; Campbell et al., 2013; Laale, 1977).

A popular approach that makes zebrafish model enormously extraordinary is “transgenesis”. Many green or red glowing transgenic zebrafish can be constructed.

Taking advantage of the extraordinary transparency of zebrafish embryos and an over-growing number of transgenic lines, multiple macrophages, neutrophils, and other leukocytes fluorescent reporter lines are raised by the researchers. Such transgenic reporter lines allow real-time visualization of the infection and immune cell behaviour (Astin et al., 2017).

A very interesting feature of zebrafish is the clear chronological separation between innate and adaptive immunity. The brief development period from fertilization to hatching takes approximately 96 hours at 26°C. Primitive macrophages differentiate in yolk sac and appear in the developing embryo within 22 - 24 hours post fertilization (hpf) as the first immune cells whereas, the neutrophils emerge around 36 hpf. Lymphocytes start to develop after 96 hpf but B cells cannot produce functional antibodies until 4 weeks post fertilization (wpf). Zebrafish larvae survive exclusively on the strength of their innate immune response for 4-6 wpf when the adaptive immune system becomes fully functional. So, the zebrafish embryos and larvae are a powerful *in vivo* model to exclusively study the role of vertebrate innate immune response independently from the influence of adaptive immune cells (Laale, 1977; Luukinen et al., 2018; Myllymäki et al., 2016; Torraca & Mostowy, 2018).

Intravital microscopy makes the model outstanding by allowing imaging on the cellular and subcellular levels in the optically transparent embryo and larva. The high optical transparency of zebrafish embryos provides an exciting opportunity to visualize and witness the steps of the development of infection, host-pathogen interactions, and to obtain the detailed morphogenetic information in a complex vertebrate model by live imaging in a non-invasive manner. Among several forms of light microscopy confocal fluorescence microscopy is used as a standard technology in most labs because of its good spatial resolution and the intrinsic capability to eliminate scattered light (Keller, 2013).

1.2.2 *M. marinum* – a close genetic relative of *M. tuberculosis*

M. tuberculosis is a difficult organism to work with because of its slow growth (takes more than 7 days to form colonies visible to the naked eye on 7H10 agar), requirement of BSL3 level containment laboratory, and health risk to laboratory personnel. The mouse, rabbit, or guinea pig models used for the study of TB are slow, expensive, require high levels of biosafety, and each one presents its own advantages and shortcomings (Tobin & Ramakrishnan, 2008).

Mycobacterium marinum is a close genetic relative of *M. tuberculosis* having optimum growth temperature of 25 – 35^{o}C. It is a natural pathogen of ectotherms including zebrafish, causes latent infection, systemic granulomatous inflammation and caseous necrosis in its natural hosts similar to the caseous granuloma and inflammation in human lungs caused by *M. tuberculosis* infection. It doesn't cause disease in immunocompetent individuals and is safe to use under BSL2 containment level. Moreover, *M. marinum* grows relatively rapidly and is genetically tractable organism. The detailed genomic comparison of *M. tuberculosis* and *M. marinum* is discussed in chapter 3. Human infections of *M. marinum* referred as fish tank granuloma or swimming pool granuloma are usually peripheral and restricted to the cutaneous lesions or extremities. *M. marinum* infection in zebrafish produces organized necrotic granulomas similar to the granulomas formed in the human lungs by *M. tuberculosis* infection (Figure 1.6).

With all the individual advantages of working with zebrafish and *M. marinum*, the model provides an unbeatable benefit of genetic tractability of both the host and the pathogen. Nowadays, *M. marinum* - zebrafish infection model is being used extensively to discover key features and mechanisms of human tuberculosis.

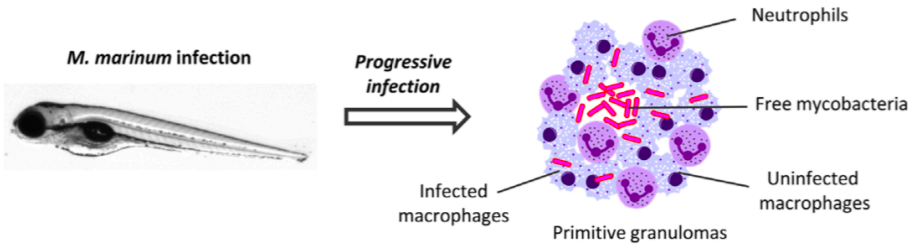


Figure 1.6 *M. marinum* infection leads to the granuloma formation in zebrafish embryos. The primary granulomas formed in fish are similar to the human granulomas in case of Mtb infection. Image taken from Myllymäki et al., 2016.

1.3 Molecular chaperones and co-chaperones

1.3.1 Definition of molecular chaperones

The term molecular chaperone was first coined by R. A. Laskey in 1978 to describe the role of a pentameric thermostable nuclear protein "nucleoplasmin" which facilitates the nucleosome assembly by inhibiting the improper interactions between DNA and histone proteins (Laskey et al., 1978). The term was used widely by R. J Ellis in 1988. Ellis defined molecular chaperones as a class of unrelated proteins that assist the proper folding (non-covalent assembly) or unfolding of many other proteins *in vivo*. Chaperones are not the constituents/components of their client proteins in their normal activity (Ellis, 1987; Ellis, 1993). Chaperones are present both in prokaryotic and eukaryotic organisms. Following are some important facts about molecular chaperones.

1) Protein assembly in the above definition refers to the proper folding of nascent polypeptides, the association of structural units to form oligomers and any deformations in the degree of folding or oligomeric association during their functions, transport of protein across membranes, or repair of heat shock damaged proteins (Lund, 2001).

2) Molecular chaperones do not include enzymes that catalyse the covalent restructurings of proteins even if they happen as a part of the protein assembly process. The mechanism of action of such enzymes does not match with the mode of action of molecular chaperones rather they function as classical enzymes (Lund, 2001).

3) Many molecular chaperones are heat-shock proteins which are induced or upregulated when cells are exposed to a certain temperature higher than their optimum growth temperature but induction upon heat-shock is not an absolute requirement for a protein to be categorized as a molecular chaperone. Some molecular chaperones are not induced by heat-shock and some heat-shock proteins are not classified as molecular chaperones (Lund, 2001).

4) Several molecular chaperones function as a complex rather than an individual protein. In this case, the term molecular chaperone is only valid to be used for the functional complex and not for the individual component proteins (Lund, 2001).

Some proteins are classified as molecular chaperones based on the *in vivo* experimental proof that they certainly assist in the assembly of other proteins. In other cases, the purified proteins demonstrated their protein folding activity *in vitro*. For genome annotation, the proteins are usually categorized as molecular chaperones based on their sequence homology with known chaperones from other organisms.

1.3.2 Importance of molecular chaperones to the cell

It had been thought that the folding of a polypeptide into an ordered functionally active structure was an essentially thermodynamically spontaneous process as suggested by C. B. Anfinsen. He demonstrated that completely reduced and denatured bovine pancreatic ribonuclease with urea was able to reorganize to an active enzyme under optimal conditions (Anfinsen et al., 1961). This is in general true for many proteins which are denatured by chaotropic agent like urea. Upon removal of urea, in majority of cases, a considerable proportion of the inactivated/denatured protein refolds to form the active protein. This shows that all the necessary information for secondary and tertiary structure of a given protein is contained in the amino acid sequence of the polypeptide. Because of this it had been assumed that in the native cell environment proteins do not need any assistance for proper folding to their active conformation (Lund, 2001). However, this concept was incorrect because of the following reasons.

In the process of *de novo* folding of nascent proteins in the cells, the proteins can start to fold before the complete synthesis of a polypeptide during the protein synthesis on ribosome. Because of this, *de novo* folding of a nascent protein as it emerges from the

ribosome exit tunnel cannot easily be compared to the refolding of a full-length protein which has been denatured in vitro. Proteins also tend to fold by hydrophobic interactions so nascent proteins or proteins being synthesized on the ribosomes can also mingle with each other through hydrophobic interactions. Therefore, it is necessary to provide protection to the growing polypeptides until the complete protein is synthesized and this role is played by some molecular chaperones (Lund, 2001).

After complete synthesis of the polypeptide from ribosomes, the in vivo folding of proteins is extremely different process than *in vitro* folding of a denatured protein. The concentration of protein in the *in vitro* experiment (as performed by Anfinsen et al., 1961) is usually low around 1.0 mg/ml. If experiments are performed using higher concentrations, then denatured proteins aggregate together before refolding back to the native conformation. In the native cell environment, proteins are at high risk of forming aggregates before they assemble to their native conformation, so some molecular chaperones prevent the proteins to aggregate (Lund, 2001).

Another important point to note is that the unfolded state of the protein is favoured than the native state in terms of free energy. Wolynes et al., 1995 represented the energy landscape of a folding protein in the form of an uneven funnel showing all the energy levels available to the protein during its folding. The unfolded protein state has highest free energy which is folded to its most stable native conformation of lower free energy by molecular chaperones using one of several routes available (Nelson Onuchic et al., 2000) (Figure 1.7). Because of this, proteins become more vulnerable to increase in temperature. A slight increase in temperature which is not sufficient to cause denaturation of all the proteins could cause unfolding and subsequent aggregation of some essential proteins which could be lethal to the cell. Molecular chaperones prevent the protein aggregation at higher temperatures and reverse the unfolded or aggregated proteins to their native state (Lund, 2001). Therefore, many molecular chaperones are also called heat shock proteins.

Many proteins travel from cytoplasm to the periplasm or across plasmid membrane and for this, proteins should be in their unfolded conformation. The proteins are protected partly by molecular chaperones until they reach their final destination. And finally chaperones also assist the assembly of protein to form homo or hetero

oligomers by protecting the complementary surfaces from binding to the incorrect partners (Lund, 2001).

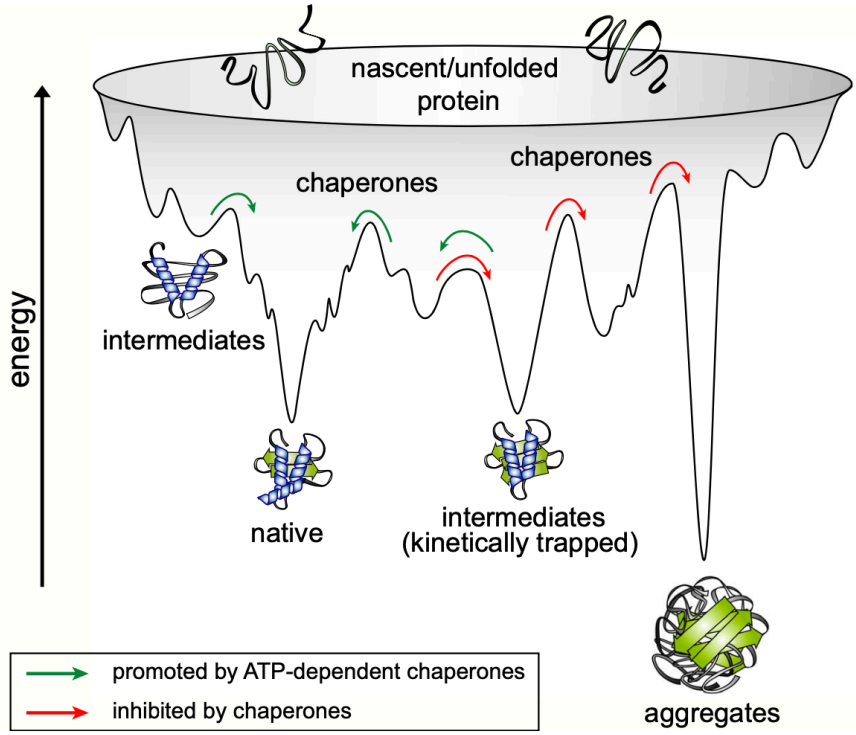


Figure 1.7 The energy landscape of protein folding shaped by molecular chaperones. Image taken from (Balchin et al., 2020). Molecular chaperones prevent the aggregation and assist the assembly of the unfolded proteins by providing kinetic assistance and separating the kinetically trapped folding intermediates from the native conformation.

The DNA sequences of molecular chaperones are observed to be greatly conserved in nature. Chaperones are present in all the three domains of cellular life i.e., Bacteria, Archaea, and Eukarya (Lindquist & Craig, 1988). High sequence homology has been observed among the families of molecular chaperones. Molecular chaperones are essential for maintaining the protein homeostasis of the cells. The study of chaperones, their interactions, and functional mechanisms is critical for understanding the cellular homeostasis and survival strategies of cells under normal and stressed conditions. For example, the study of the role of chaperones in the survival mechanisms adapted by *M. tuberculosis* inside the extremely stressed environment of granulomas would be helpful in understanding the survival strategies of the bacilli and could provide a novel molecular marker for the disease pathogenesis.

1.3.3 Classes of molecular chaperones

The classification of molecular chaperones is done based on their molecular weight, mechanism of action, the client proteins (substrates), and phylogenetic relationships. Several different classes of chaperones have been characterised which function either independently or in combination with other chaperones to assist the assembly of their client proteins in prokaryotes and eukaryotes (Table 1.1). Chaperones can be distinguished based on their ATP dependent or ATP independent mode of action. Some classes of chaperones such as Hsp60 and Hsp70 require ATP hydrolysis for binding and release of their client proteins (Goloubinoff et al., 2018). A few chaperones can work in ATP independent fashion such as sHsps and Tig or Trigger factor (Hoffmann et al., 2010; Janowska et al., 2019). Some chaperones require a co-chaperone for performing their chaperone activity. For example, Hsp60 (GroEL) requires a co-chaperone (GroES) for performing its function of protein folding.

Chaperones are also classified based on their functions and mode of action. Some chaperones are classified as foldases which actively participate in the assembly of unfolded proteins usually in ATP-dependent fashion e.g., Hsp60 and Hsp70 chaperones. Other chaperones can be classified as holdases because they actively and efficiently bind to the folding intermediates to prevent them to form aggregates. The example of such chaperones are sHsps and Hsp40. A third class of chaperones are disaggregases which are specialized in disaggregating the aggregated proteins in a non-destructive way e.g., Hsp100 (Barends et al., 2010; Kim et al., 2013; Richter et al., 2010). A useful way of classification of molecular chaperones is the classification based on their sequence homology which also provides some information about the structure and functions of a protein. Based on sequence similarity proteins are classified in different categories as shown in Table 1.1.

Sr no.	Chaperone family	Example in bacteria	Example in Archaea	Example in Mitochondria	Example in chloroplast	Example in eukarytic cytosol	Example in ER	Example in *E. coli*	Function in *E. coli*
1	Hsp100	ClpA, ClpB, ClpX	Absent	HSP78	ClpC, ClpB	HSP104, Not found in cytosol of animals	Absent	ECK0873 (ClpA), ECK0873 (ClpB), ECK0432 (ClpX)	ATP-dependent specificity component of the ClpAP protease. It directs the protease to specific substrates. It has unfoldase. activity. The primary function of the ClpA-ClpP complex appears to be the degradation of unfolded or abnormal proteins.
2	Hsp90	HtpG	Absent	TRAP1	HSP90	HSP90, HSP82, HSP83	GRP94	ECK0467 (HtpG)	Molecular chaperone. Has ATPase activity.
3	Hsp70	DnaK	HSP70 (Absent from many Archaea)	SSC1, Grp75	Absent	HSP70, HsC70, SSB, SSA	BiP	ECK0014 (Dnak)	1) Essential role in the initiation of phage lambda DNA replication. 2) Involved in chromosomal DNA replication. 3) Actively participates in the response to hyperosmotic shock.
4	Hsp60	GroEL, Hsc66	Thermosome, TF55, CCT	HSP60	Rubisco sub-unit binding protein	TRiC, CCT	Absent	ECK4137 (GroEL)	Prevents misfolding and promotes the refolding and proper assembly of unfolded polypeptides generated under stress conditions.
5	Hsp40	DnaJ, CbpA, DjlA	HSP40 (absent from many Archaea)	Ydj1	Cdj1	HSP40, Sis1, Zuo1	Sec63	ECK0015 (DnaJ)	1) Participates actively in the response to stress e.g., heat shock. 2) Preventing the aggregation of stress-denatured proteins. 3) Also work in an autonomous, DnaK-independent fashion. 4) Interacts with DnaK and GrpE (replication of phage lambda).
6	sHsps	IbpA, IbpB	HSP16.5	HSP22	HSP21	HSP26	Absent	ECK3679 (IbpA), ECK3678 (IbpB)	1) Associates with aggregated proteins. 2) stabilize and protect them from irreversible denaturation and extensive proteolysis during heat shock and oxidative stress. 3) Proteins bound to the IbpAB complex are efficiently refolded.
7	Trigger factor	Tig	Absent	Absent	Present	Absent	Absent	ECK0430 (Tig)	1) Involved in protein export. 2) Acts as a chaperone by maintaining the newly synthesized secretory and non-secretory proteins in an open conformation. 3) Binds to nascent polypeptide chains via ribosomal protein L23.
8	Co-chaperone	GrpE	Absent	Absent	Absent	Absent	Absent	ECK2610 (GrpE),	1) Participates actively in the response to hyperosmotic and heat shock by preventing the aggregation of stress-denatured proteins. 2) Works in in association with DnaK and GrpE. 3) Nucleotide exchange factor for DnaK, function thermosensor.
9	Co-chaperonin	GroES	Absent	Absent	Absent	Absent	Absent	ECK4136 (GroES)	Binds to Cpn60 in the presence of Mg-ATP and suppresses the ATPase activity of the latter.

Table 1.1 Different classes of molecular chaperones and their examples in bacteria, archaea, and eukaryotic cells.
ER = Endoplasmic reticulum. The examples given for each class of chaperone family are not the only examples but are a few known chaperones belonging to that family (Bross et al., 2013; Karunanayake & Page, 2021; Kumar et al., 2015; Lund, 2001; Ranford et al., 2000).

1.3.4 Examples of molecular chaperones in *M. marinum* and *M. tuberculosis*

The examples of different chaperones from *M. marinum* and *M. tuberculosis* are given in Table 1.2. Both the mycobacteria have two Hsp60 proteins, one is called Cpn60.1 and the other is called Cpn60.2. The detail and nomenclature of these proteins is given in section 1.4.1.

Sr no.	Chaperone family	Example in *E. coli*	*M. marinum*	*M. tuberculosis*
1	Hsp100	ECK0873 (ClpA), ECK0873 (ClpB), ECK0432 (ClpX)	MM1218R_RS03020 (ClpB)	Rv0384c (ClpB)
2	Hsp90	ECK0467 (HtpG)	MM1218R_RS17950 (HtpG)	Rv2299c (HtpG)
3	Hsp70	ECK0014 (Dnak)	MM1218R_RS02960 (DnaK)	Rv0350 (DnaK)
4	Hsp60	ECK4137 (GroEL)	MM1218R_RS05445 (Cpn60.1) MM1218R_RS03575 (Cpn60.2)	Rv3417c (Cpn60.1) Rv0440 (Cpn60.2)
5	Hsp40	ECK0015 (DnaJ)	MM1218R_RS02970 (DnaJ1) MM1218R_RS18770 (DnaJ2)	Rv0352 (DnaJ1) Rv2373c (DnaJ2)
6	sHsps	ECK3679 (IbpA), ECK3678 (IbpB)	MM1218R_RS02420 (Hsp)	Rv0251c (Hsp20, Hsp, Acr2)
7	Trigger factor	ECK0430 (Tig)	MM1218R_RS19350 (Tig)	Rv2462c (Tig)
8	Co-chaperone	ECK2610 (GrpE), ECK4136 (GroES)	MM1218R_RS02965 (GrpE) MM1218R_RS05440 (GroES)	Rv0351 (GrpE) Rv3418c (GroES)

Table 1.2 Different classes of molecular chaperones and their examples in *E. coli*, *M. marinum*, and *M. tuberculosis.*

1.3.5 Role of different classes of molecular chaperones in protein homeostasis

Different classes of chaperones play different roles in quality control of the cell proteome. Newly synthesized proteins from ribosomes are either released in unfolded/non-native state or in folded state because of ribosome associated chaperones (Richter et al., 2010) (Figure 1.8 - 1,2). The non-native, unfolded or misfolded proteins released from the ribosome, unfolded due to stress, or disaggregated proteins are assembled to their native conformation either spontaneously or by the assistance of molecular chaperones such as Hsp70 (DnaK), Hsp60 (GroEL), and Hsp90 (Papadimitriou et al., 2016) (Figure 1.8 - 3).

Nascent proteins or denatured proteins (unfolded due to stress) are inclined to form aggregates which could be lethal to the cells; some classes of chaperones are specialized to prevent the aggregates formation such as Hsp100 (Clp), small heat shock proteins sHsps, Hsp70 (DnaK), and Hsp90 (Barends et al., 2010; Henderson et al., 2006) (Figure 1.8 - 4). These chaperones also disaggregate the aggregated proteins thus reverse the aggregation process and return the proteins to the pool of unfolded proteins. In cases where unfolded proteins cannot be assembled back to the native state, they are degraded by Hsp100 (Clp) (Franke et al., 2017; Li & Sha, 2003) (Figure 1.8 - 5).

A generalized model of chaperones assisted protein assembly is shown in Figure 1.8 -A. All nascent proteins do not require chaperones for their assembly even if require chaperones to stop aggregation. Many small proteins after being released from ribosomes spontaneously fold themselves without any assistance. Trigger factor (Tig) and Hsp70 chaperones hold the polypeptide newly released from ribosomes and allow them to fold without any disturbance from the surrounding environment. Tig proteins usually bind directly with nascent polypeptides allowing 60 – 80 % of small size proteins to fold into their native state spontaneously (Hoffmann et al., 2006). The proteins which cannot fold in this way are presented to the Hsp70 complex (Deuerling et al., 1999; Thulasiraman, 1999). Some proteins demand slow-folding process and are more prone to aggregation. Such proteins are assembled by chaperone Hsp60 which provides an enclosed chamber for safe folding into native state (Houry et al., 1999).

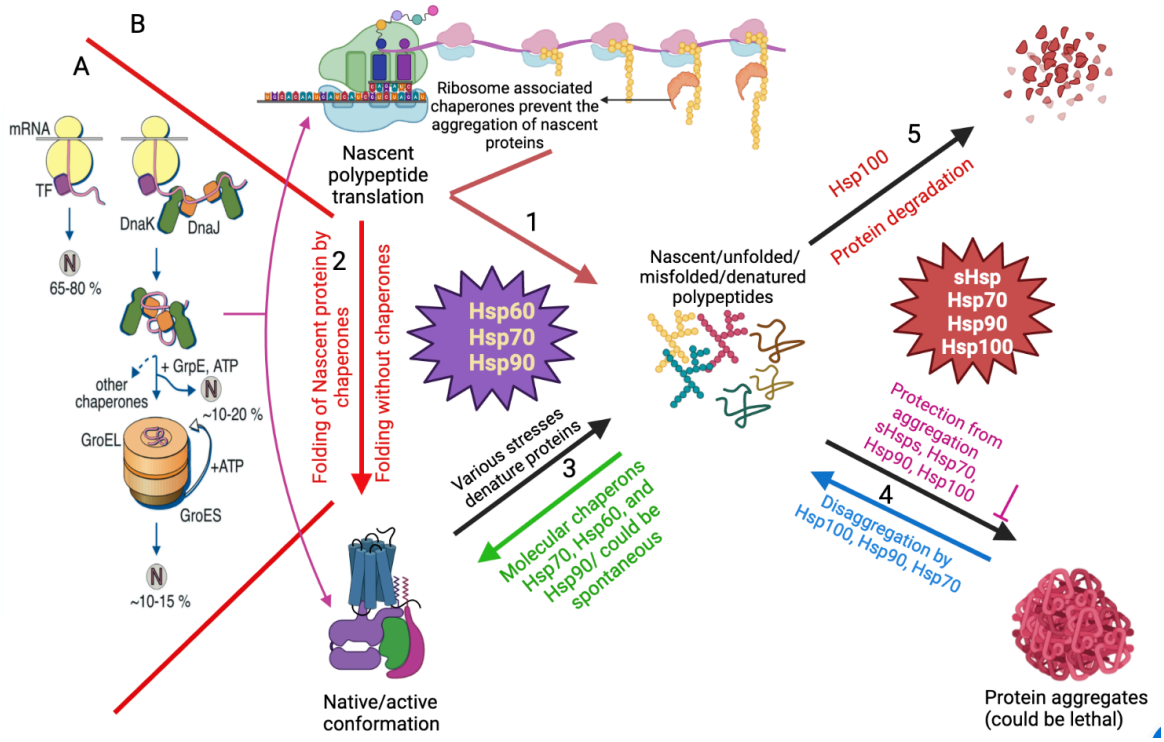


Figure 1.8 Chaperone assisted protein folding and protein homeostasis.
A) General model of chaperone assisted protein folding of nascent protein. B) Schematic illustration showing the role of different chaperonin families in the protein homeostasis of a cell. Figure adapted from (Hartl & Hayer-Hartl, 2002).

Several studies have demonstrated that the GroE proteins (GroES and GroEL) and DnaK/DnaJ/GrpE proteins perform the major chaperone activities of assisting the assembly of some proteins at normal and stressed conditions (Ang & Georgopoulos, 1989; Gragerov et al., 1992) (Table 1.1). These two sets of chaperones are considered to work as two major chaperone complexes in bacteria therefore are discussed in detail in the next two sections.

In general, for protein folding molecular chaperones bind to the non-native proteins. In non-native state the hydrophobic amino acids of the protein are exposed and are recognized by the molecular chaperones. Molecular chaperones do not have any information about the structural conformation of the proteins they fold rather they prevent the pre-mature folding and unwanted interactions of the client proteins. This is accomplished by controlled binding and release of the client proteins which is dependent on ATP hydrolysis (Figure 1.9).

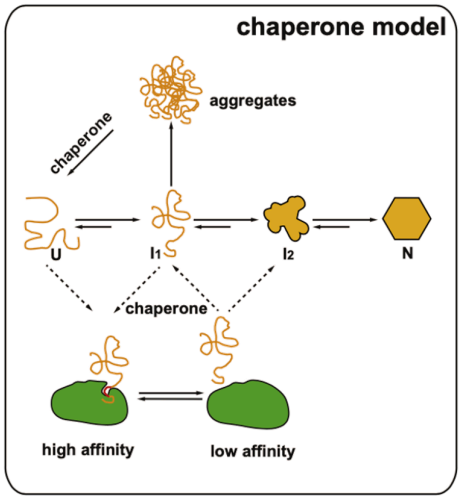


Figure 1.9 A generalized model of mechanism of chaperone assisted folding.

Chaperones bind proteins in unfolded state (U) and release them in folded state (N). During conversion from unfolded to folded state, structural intermediates can also be formed. A shift from high affinity to low affinity state is ATP dependent. Image taken from (Richter et al., 2010).

1.3.6 DnaK/DnaJ/GrpE proteins chaperone complex

Hsp70 is one of the most conserved chaperones. In prokaryotes, the protein DnaK shares 60% sequence identity with the Hsp70 protein of eukaryotes. DnaK/DnaJ complex is second major chaperone machine of bacterial cells. In *E. coli* genome, DnaK is a 638 amino acid containing protein. Its N-terminal has a nucleotide binding domain which also possess ATPase activity and its C-terminal contains a substrate binding domain. Both domains are connected by a highly conserved linker of 6 amino acids (Castanié-Cornet et al., 2014; Karunanayake & Page, 2021) (Figure 1.10).

The two co-chaperones of DnaK in *E. coli* are DnaJ and GrpE. DnaJ is a 376 amino acid protein, and its gene is present downstream to *dnak* in *E. coli*. GrpE consists of only 197 amino acids and its gene is found separately on *E. coli* chromosome (Figure 1.11). DnaJ has a highly conserved domain called J-domain. This domain is present in many proteins that interact with Hsp70 homologous protein.

DnaK/DnaJ binds transiently with unfolded, nascent polypeptides or partially folded proteins. DnaK/DnaJ are known to bind with proteins being synthesized on the ribosome and protect them from misfolding and aggregation before they are completely translated. At heat shock condition, DnaK plays a role of reactivation of heat denatured proteins by working in combination with ClpB chaperone. Therefore, we can

divide the role of DnaK chaperones in two categories. First, they work as holdases and prevent the incorrect or pre-mature folding of the translating polypeptides by preventing their aggregation. Second, in combination with ClpB they work as refolders by refolding the aggregated proteins and bringing them back to the native state (Lund, 2001).

The main points of the current proposed mechanism of reaction cycle of DnaK are as follows.

1) The client proteins either bind directly with DnaK-ATP or its largest co-factor DnaJ binds to the unfolded proteins and delivers them to DnaK.

2) The ATP is hydrolysed and DnaK-ADP-substrate-DnaJ is formed. This complex is more stable because the post-hydrolysis ADP state of DnaK has higher affinity for peptide binding.

3) GrpE protein works as a nucleotide exchange factor which replaces ADP with ATP. This way DnaK is converted back to the low-affinity state hence, DnaJ and the substrate are released (Lund, 2001; Richter et al., 2010) (Figure 1.12).

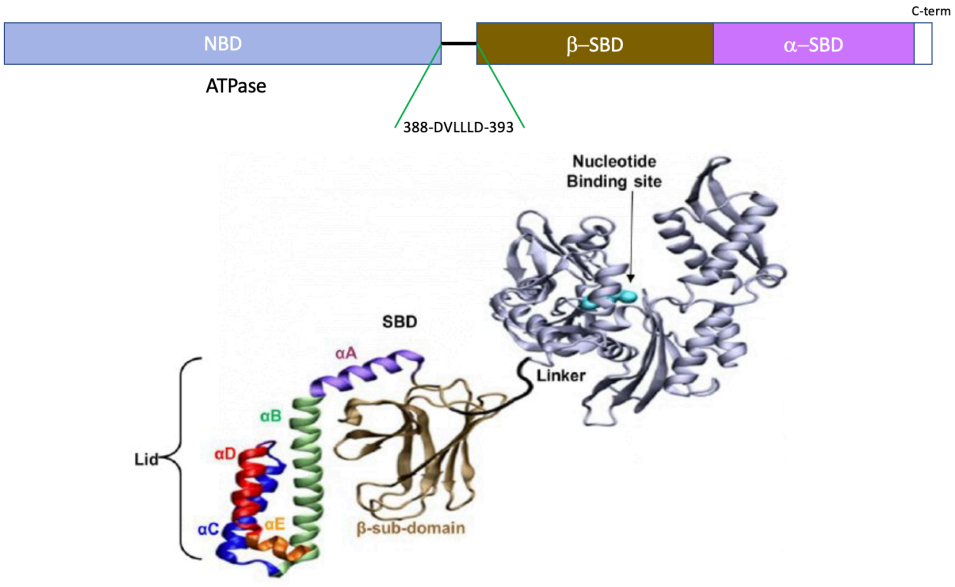


Figure 1.10 The domain regions and crystal structure of DnaK protein of *E. coli*.
Nucleotide binding domain (NBD) and substrate binding domain (SBD) are shown. Both domains are linked by a 6 amino acid linker DVLLLD (Castanié-Cornet et al., 2014; Karunanayake & Page, 2021). The crystal model is taken from (Azoulay et al., 2013).

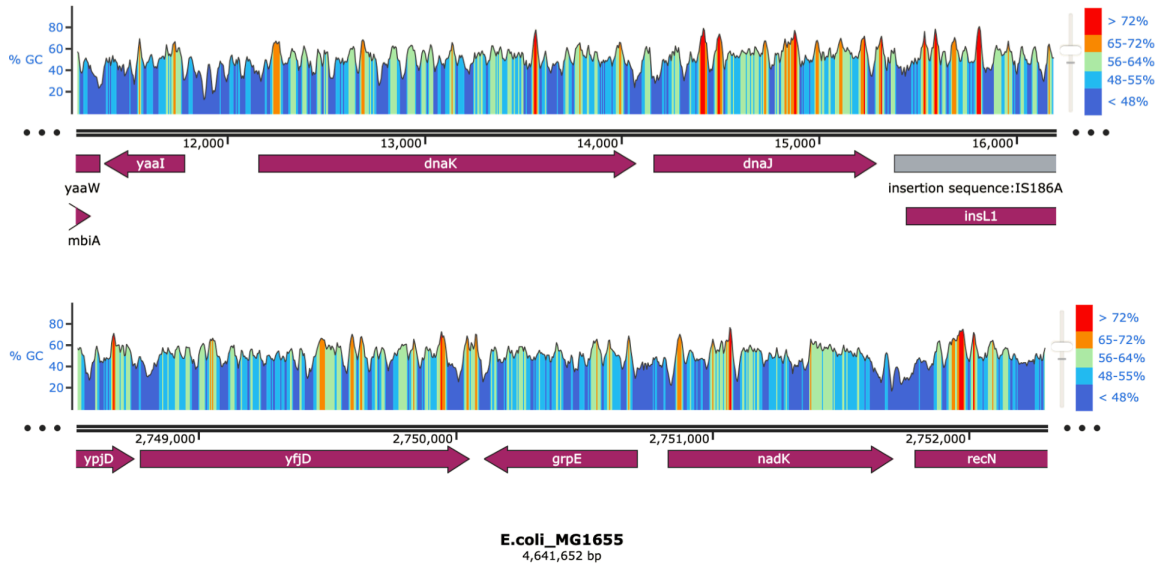


Figure 1.11 Genomic organization of *dnaK* , *dnaJ*, and *grpE* genes in *E. coli* genome.
dnaK (51% GC), *dnaJ* (55% GC) and *grpE* (50% GC). There is only 88 bps gap between *dnaK* and *dnaJ* genes whereas *grpE* is present separately on chromosome.

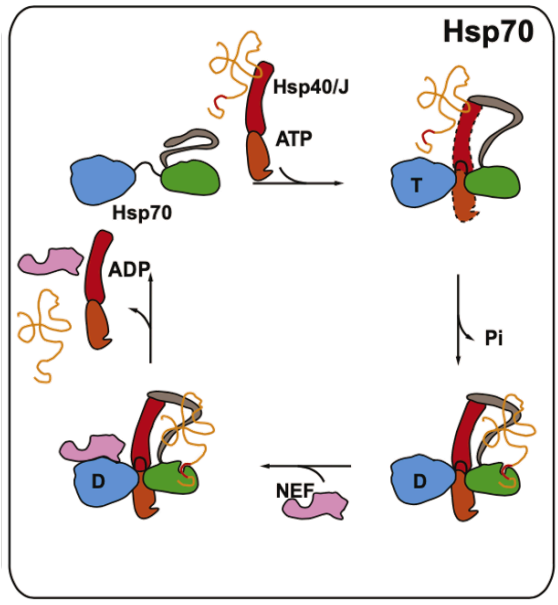


Figure 1.12 DnaK/DnaJ mechanism of reaction.
Image taken from (Richter et al., 2010).

1.3.7 Chaperonins – GroEL/GroES complex

The term chaperonin was first used by Ellis to refer only one class of molecular chaperones which is GroEL of *E. coli* and its homologous proteins (Hemmingsen et al., 1988). GroEL (548 amino acids) is Hsp60 chaperone in *E. coli*. It works in combination

with a co-chaperonin GroES (97 residues). The gene encoding GroES (*groES)* is present upstream to *groEL* in *E. coli* (Figure 1.13). GroEL protein demonstrate a high sequence and function conservation across different kingdoms of life. The most strictly conserved residues were found to be related to the ATP/ADP and magnesium ions binding sites. A significant conservation was also found in the hydrophobic amino acids which are involved in substrate binding (Brocchieri & Karlin, 2008; Lund, 2001). In *E. coli* the chaperonin GroEL and its co-chaperonin GroES are essential for the survival of the cell.

The GroES-GroEL complex can also perform its protein folding activity *in vitro*. This was first demonstrated by Goloubinoff and his colleagues in 1989 by *in vitro* refolding of dimeric ribulose bisphosphate carboxylase (Rubisco) using GroES-GroEL and Mg-ATP of *E. coli*. Only a fraction of unfolded Rubisco refolded into its active form in vitro without the addition of GroES-GroEL. Upon addition of GroES-GroEL and Mg-ATP to the refolding buffer, nearly 100% yield of active Rubisco was obtained (Goloubinoff et al., 1989). Nearly 5% of *E. coli* cytoplasmic proteins obligately require GroES-GroEL assistance for proper folding into their native conformation (Kudryavtseva et al., 2021). Some of the obligate client proteins of GroES-GroEL are essential for survival of the cell. Therefore, the *groES* or *groEL* deletion mutant of *E. coli* cannot sustain growth unless a complementing chaperonin is provided.

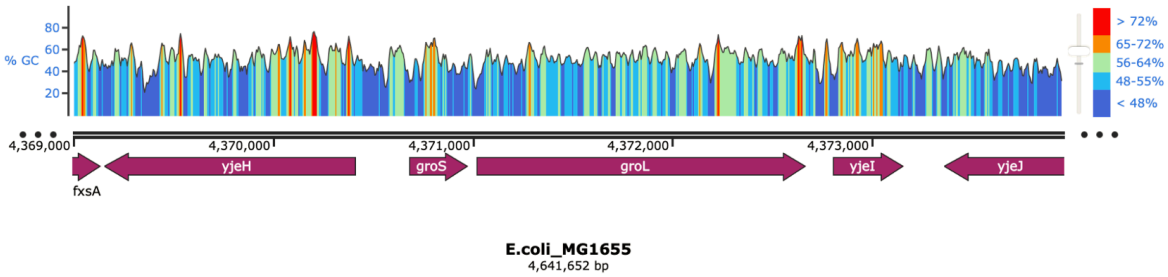


Figure 1.13 Genomic organization of GroES-GroEL complex of *E. coli*.
The gene for chaperonin GroEL (*groL,* 53% GC*)* and its co-chaperonin GroES (*groS,* 49% GC*)* are found in a single operon in *E. coli*.

1.3.7.1Structure of GroEL and GroES

The *E. coli* chaperonin GroEL forms a tetradecamer of two rings of seven identical subunits which are stacked back-to-back (cis and trans rings, Figure 1.14-C). The

molecular weight of each subunit (monomer) is ~58 kDa. Each monomer comprises three domains consisting of discontinuous sequence: equatorial domain, intermediate domain, and apical domain. Equatorial domain is an ATP binding domain which consists of residues 6 – 133 and 409 – 523. The intermediate domain which forms the hinge consists of residues 134 – 190 and 377 – 408 and apical domain consists of 191 – 376 amino acids (Figure 1.14-A). The apical domain binds to the client proteins and GroES (Figure 1.14-B). The 23 amino acids of the C-terminal of each monomer bulge from the equatorial domain into the ring cavity (Hayer-Hartl et al., 2016).

All the intra and inter ring contacts of the tetradecamer are provided by equatorial domains. The two heptameric rings stack in a cis trans manner. The apical domains of the heptamer form the mouth or opening of the heptameric ring whereas, the intermediate domain of each monomer connects the apical domain to the equatorial domain and transmit conformational change to the ring. The two helices of the apical domain, helix H and helix I, labelled in the Figure 1.14-B expose the hydrophobic amino acids to the centre of the ring forming a surface for binding the client proteins. GroES forms a dome shaped heptameric lid of ~10 kDa monomers. Each monomer folds to form a single domain of nine β strands (Xu et al., 1997a).

The GroESL complex protein folding machine is a fundamental requirement of cells under all growth and temperature conditions but becomes particularly more important under stress conditions such as high temperature causing denaturation of cellular proteins. It probably plays the same role of protein folding at normal and heat-shock conditions. The client proteins remain the same, but proportion of each client protein is increased at high temperature (Lund, 2001).

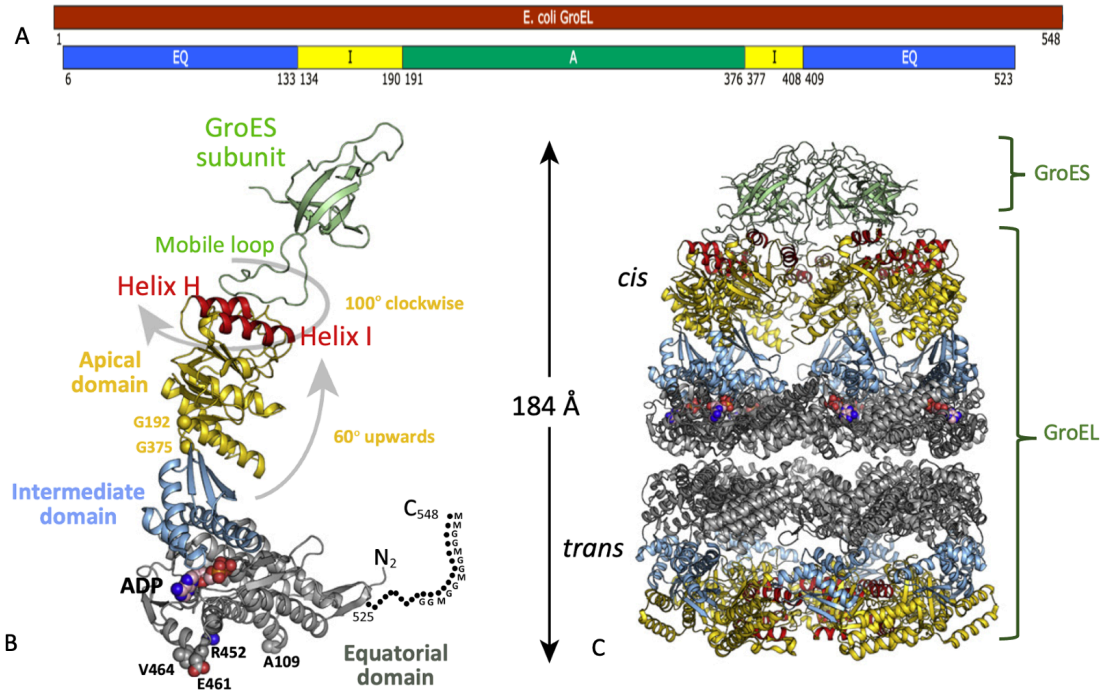


Figure 1.14 Domain regions and crystal structure of GroES and GroEL.
Domain regions of GroEL. B) Crystal structure of a monomer of GroEL bound with the monomer of GroES. Apical domain (yellow), intermediate domain (blue), and equatorial domain (gray). C) The tetradecamer structure of GroEL with the GroES lid. Image taken from (M. Hayer-Hartl et al., 2016).

1.3.7.2GroES-GroEL protein folding mechanism

Each heptameric ring of GroEL has a central cavity which is large enough to contain a 60 kDa protein. GroES forms the lid, it can bind with both ends of the two heptamers but at a time it sits only on one end of the GroES-GroEL complex. The two cavities of the tetradecamer are not interconnected. The different stages of the GroES-GroEL cycle of protein folding are as follows (Lund, 2001).

1) The best *in vitro* acceptor state of the GroEL-GroES complex to start the cycle is ADP molecules bound to one of the heptameric ring where GroES is bound. This complex does not exist *in vivo* but is a convenient state to start the reaction cycle. In the first step of a new cycle the client protein binds to one of the heptameric rings of the tetradecamer that is not connected with GroES (cis ring). This binding of client protein to the apical domain occurs by hydrophobic interactions (Figure 1.15, Step 1) (Lund, 2001; Ranson et al., 2006; Weissman et al., 1996).

2) After substrate binding, seven ATP molecules bind to the equatorial domains of the cis ring subunits followed by the binding of the GroES the cis ring. This causes loss of ADP and GroES from trans ring. The bound client protein is moved to the cavity of the heptameric ring. At this stage conformational change of the heptameric cavity and the transformation of inner surfaces of the ring from hydrophobic to hydrophilic takes place (Figure 1.15, Step 2) (Lund, 2001).

3) The ATP in the cis ring is slowly hydrolysed. The binding of GroES and ATP converts the cavity from peptide binding state to the folding active state (Figure 1.16). The client protein is now enclosed in the cavity where it is free to fold in the absence of unwanted interactions of the cytosolic environment. The protein folds into its native conformation until the bound ATP molecules are hydrolysed (~10 s). After ATP hydrolysis, the GroEL-GroES system look like the acceptor state we started with, but it contains the client protein inside its cavity. This could be the best *in vivo* acceptor state of the ring (Figure 1.15, Step 3) (Lund, 2001; Weissman et al., 1996; Xu et al., 1997a).

4) A new substrate and seven ATP molecules bind to the trans ring (the cis ring for next cycle) (Figure 1.15, Step 4) (Lund, 2001).

5) The binding of ATP and substrate to trans ring leads to the dissociation of GroES, discharge of ADP, and substrate protein from the cis ring. This permits the binding of GroES to the trans ring (which is the new cis ring for next cycle) (Figure 1.15, Step 5). The released substrate could be in folded state, or it may require more cycles of GroEL binding to reach the final native conformation. The complex is now ready to enter the next cycle. This is an inverted copy of the step 2 complex (Lund, 2001).

In general, it is assumed that the GroESL model of *E. coli* is the best representative model for the chaperonins of other bacteria. In many studies it has been observed that the heterologous GroE proteins from other organisms can in general complement the loss of any of the GroE proteins of *E. coli* (Lund, 2001). This shows that different GroE chaperonins are able to bind to and fold the same essential client proteins.

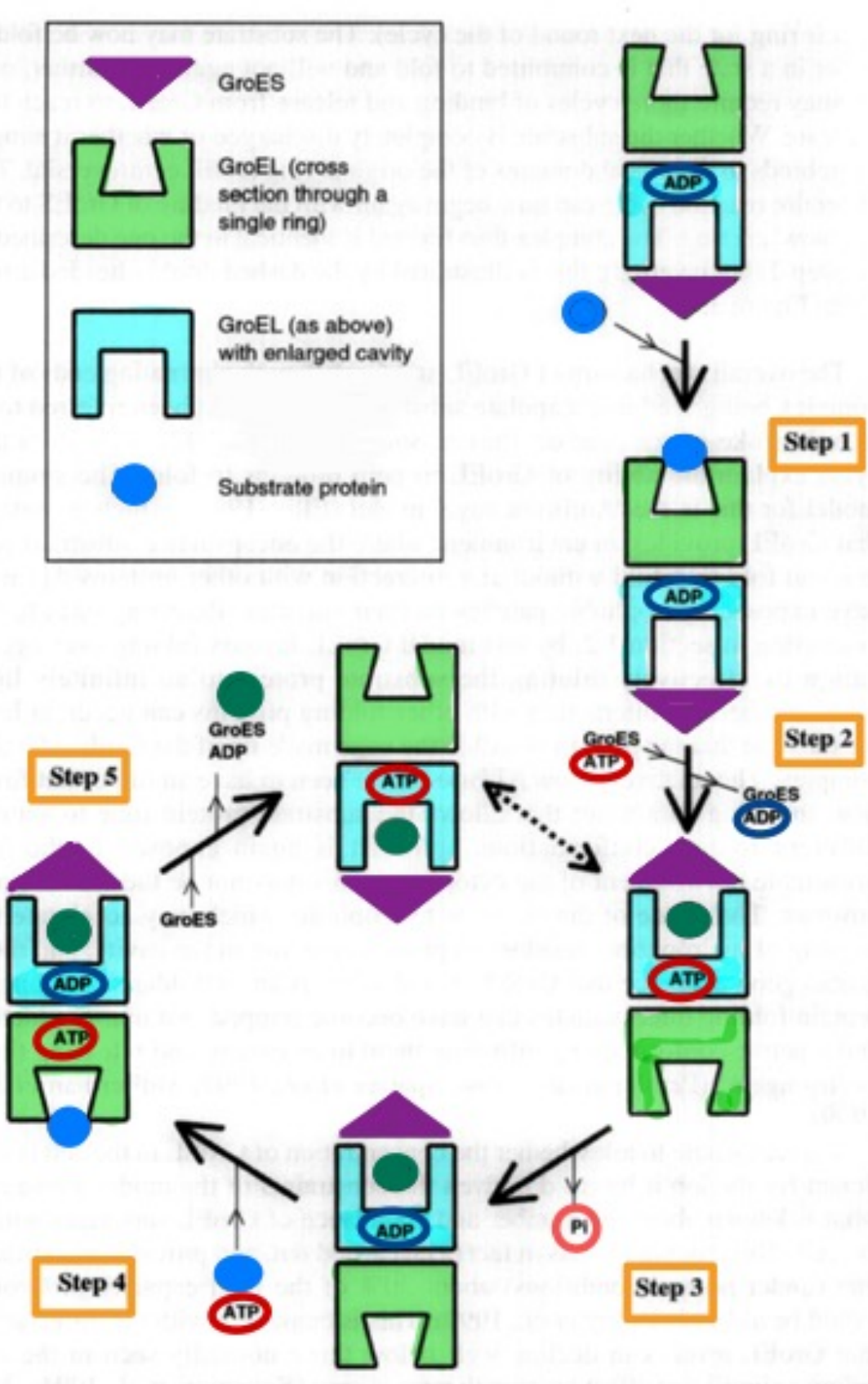


Figure 1.15 A model for GroES-GroEL protein folding mechanism. Figure adapted from (Lund, 2001).

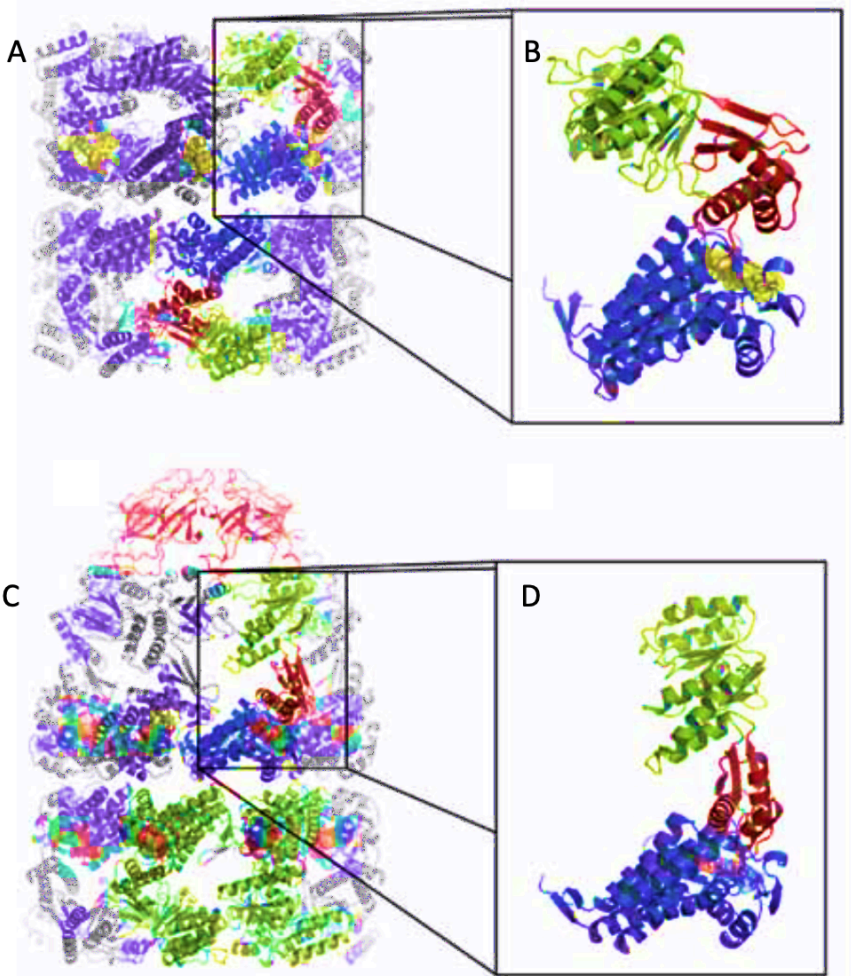


Figure 1.16 Conformational change in the structure of GroEL complex upon substrate binding. A) The two heptameric rings of GroEL stack back-to-back. Two subunits (one in each heptamer) are coloured. C) GroES-GroEL-ADP7 complex. B & D) One monomer showing the conformation change of a single subunit when not bound to the substrate and GroES (B) vs ADP bound (A) forms. Figure is taken from (Skjærven et al., 2011).

1.3.7.3 Moonlighting activities of GroEL

In *E. coli*, GroEL is always found in the cytoplasm of the cell. However, in some bacteria, GroEL homologs are observed to be associate to the surface of the pathogenic bacterial species such as *Legionella pneumophila* (Garduño et al., 2011; Valenzuela-Valderas et al., 2021) and *Helicobacter pylori* (Vanet & Labigne, 1998). The function of the secreted GroEL homologue of *H. pylori* is not well known but its binding with host iron was reported by Senkovich et al., 2010. This could be advantageous for the growth of the *H. pylori* during infection because *H. pylori* requires iron binding for its growth. Iron ions are essential elements of tricarboxylic acid cycle, DNA and amino acids synthesis, and many other critical cellular processes.

The GroEL homologous secreted outside of the cell are also observed to be associated with the pathogenesis. Paju et al., 2000 reported the presence of GroEL homologue in the membrane fractions of *Aggregatibacter actinomycetemcomitans* strain. The strains expressing more GroEL homologous protein in the membrane fraction induced the proliferation of infected epithelial cells more strongly compared to those which weakly expressed the protein. This propose that the GroEL homologous proteins might be involved in proliferation of infection. These observations suggest closer investigation of bacterial chaperonins and their homologous protein specifically of pathogenic species.

GroEL proteins of *Lactobacillus johnsonii, Mycoplasma pneumoniae*, and *Salmonella enteric*a were also observed to be present on the surface of the cells. Gabriela reported that GroEL of *L. johnsonii* was found on the cell surface and it might be involved in aggregation of the gastric pathogen, *H. pylori* (Bergonzelli et al., 2006). The GroEL homologue of *Salmonella typhimurium* was found to be relocated to the surface and involved in binding to the mucus present in the intestine of the infected host (Ensgraber & Loos, 1992). In another study, the surface associated GroEL of *M. pneumoniae* was observed to play some role in pathogenesis by interacting with several human enzymes and proteins such as fibronectin, plasminogen, and lactoferrin (Hagemann et al., 2017).

GroEL was also found to act as a toxin for some insects. For example, the GroEL excreted from *Enterobacter aerogenes* paralyzed cockroaches, but it could not cause the same effect in mice. It was speculated that the effect of toxin might be species specific (Yoshida et al., 2001). It is still to be discovered how GroEL homologs are secreted out of the cell and what mechanism switches the protein to act as a virulence factor from its normal chaperone activity.

Several studies have shown the moonlighting role of bacterial chaperonin proteins to be involved in activities other than protein folding such as bacterial adhesion, proliferation of infection, and immune signalling. It was also discussed in section 1.1.2 that a mycobacterial Cpn60.1 (GroEL homologue) was observed to have important role in pathogenesis and immune signalling (Figure 1.5). We are interested in better understanding of the importance and role of chaperonins in mycobacterial infection.

In the following section the general information available in literature about the chaperonins of mycobacterial species is discussed. Most of the bacteria are found to have only one chaperonin genes and a co-chaperonin but some bacterial genomes are found to have more than one chaperonin genes; mycobacterial species are one of them.

1.4 Multiple chaperonins

With the exception of a small number of mycoplasmas, *cpn60* and *cpn10* genes are present in all bacteria usually in a single operon. About 30% of bacterial genomes are found to have two or more chaperonin genes (Lund, 2009). Actinobacteria are the first group of bacteria found to have more than one chaperonin genes (Kumar et al., 2015). Most of the Actinobacteria are found to have two, three or four *cpn60* genes and only one *cpn10* gene. s

1.4.1 Nomenclature of chaperonins

The first chaperonin proteins discovered were two GroE proteins of *E. coli* i.e., GroES and GroEL. GroEL and GroES names should be used only for the chaperonins of *E. coli* because they were discovered initially as genes which when mutated prevented the phage growth (Georgopoulos et al., 1972).The GroEL homologues were discovered in many bacteria before the full understanding of the function of GroEL. Those homologues were usually called Hsp60 or Hsp65 proteins. These proteins were also referred as “common antigens” because of their immunogenic nature. Coates et al., 1993 proposed a unified nomenclature for referring all the chaperonin proteins as Cpn proteins. Cpn proteins homologues of GroEL and GroES should be called Cpn60 and Cpn10 respectively. However, in published literature or genome annotations different names are used e.g., GroEL are also called GroL, Hsp60, or Hsp65. In this thesis the term Cpn is used for multiple chaperonin proteins in different organisms. Where two or three GroEL homologous chaperonins are present in one organism, they are called Cpn60.1, Cpn60.2, and Cpn60.3. The phylogenetic analysis of the chaperonin genes is discussed in Chapter 3 in detail. Briefly the *cpn60* genes in *Actinobacteria* fall in two distinct clusters, one is referred to as *cpn60.1* that is always found in a single operon with *cpn10* and is also missing in a few species of *Actinobacteria*. The second gene is *cpn60.2* which is essential, does not have an adjacent *cpn10* genes and always present in all known *Actinobacteria*(Lund, 2009).

1.4.2 Chaperonins of mycobacteria

All the mycobacteria are found to have at least two copies of chaperonin genes. *M. tuberculosis* and *M. marinum* have two chaperonin genes. One is *cpn60.1* in a single operon with *cpn10* and other is *cpn60.2* found separately on chromosome. *M. smegmatis* has three copies of *cpn60* genes (Figure 1.17 – Figure 1.18). The two distinct clusters of *cpn60.1* and *cpn60.2* in phylogenetic analysis reveal the divergence of function (Goyal et al., 2006).

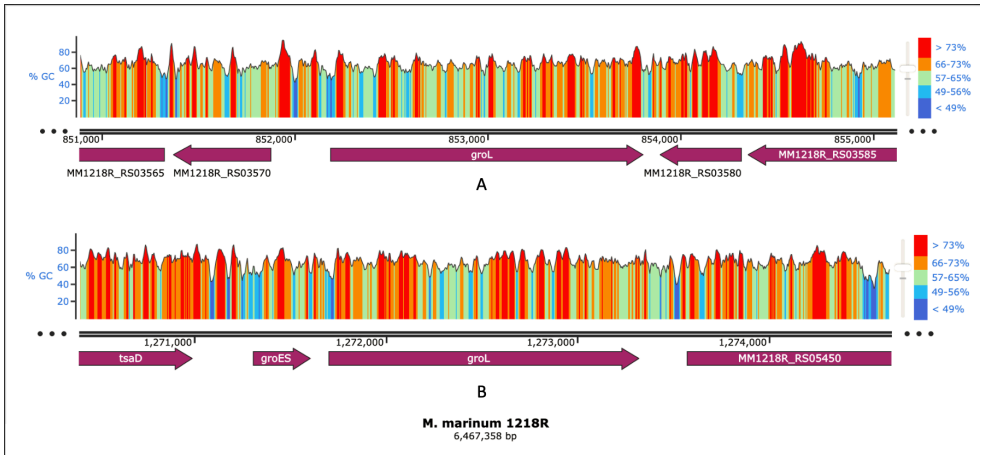


Figure 1.17 The genomic organization of cpn60.2 (A) and cpn60.1 (B) genes of *M. marinum* 1218R.
cpn60.2 (66% GC) is found separately on chromosome whereas *cpn10* (64% GC) and *cpn60.1* (68% GC) are present in a single operon.

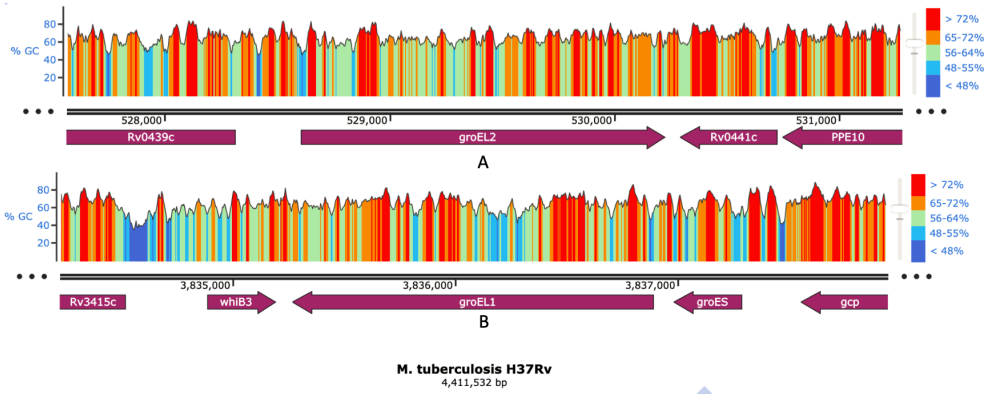


Figure 1.18 The genomic organization of cpn60.2 (A) and cpn60.1 (B) genes of *M. tuberculosis* H37Rv.
cpn60.2 (65% GC) is found separately on chromosome whereas *cpn10* (63% GC) and *cpn60.1* (65% GC) are present in a single operon.

The detailed introduction, phylogenetic and sequence alignment of *cpn60* genes of *M. tuberculosis* and *M. marinum* are discussed in chapter 3.

1.4.3 Oligomeric state

As explained in section 1.3.7 the *E. coli* GroEL has a well characterized tetradecamer structure which is essential to perform its chaperone activity and GroES forms the lid which encloses the heptameric ring cavity within which the proteins are given a separate environment for folding to their native conformation. Unlike GroEL of *E. coli*, both the Cpn60 proteins of *M. tuberculosis* are found to form lower order oligomers (Qamra et al., 2004). The oligomer formation was also tested in the presence of a variety of buffer conditions such as ATP, GroES, magnesium and potassium ions, ammonium sulphate and Cpn10. The formation of higher order oligomers was not observed in any of the tested buffer conditions.

The chaperonins of *M. tuberculosis* were also found to have reduced ATPase activity compared to GroEL (Qamra et al., 2004). Qamra & Mande, 2004 were able to crystalize the MtbCpn60.2 as a dimer which further supported the hypothesis that Cpn proteins of mycobacteria are usually found in lower order oligomeric forms. Contradictory to these findings was the complementation studies of Hu et al., 2008 where Cpn60.2 of *M. tuberculosis* was able to complement the loss of GroEL in *E. coli* in either combination with Cpn10 or GroES.

The details of our current understanding about complementation studies of different mycobacterial chaperonins are given in chapter 6. As discussed earlier that the presence of a tetradecameric double ring is essential for a chaperonin to perform its protein folding activity and the ATP binding and heptameric GroES are also required for the protein folding process. The complementation ability of Cpn60.2 of *M. tuberculosis* for the loss of *E. coli* GroEL suggests that Mtb Cpn60.2 forms higher order oligomers *in vivo*. We have investigated this further for the chaperonins of *M. marinum* in chapter 6.

1.4.4 Phosphorylation

Phosphorylation is an important post-translational modification of proteins which is observed to be important for sensing the environmental conditions. It plays important roles in regulation of several processes of the cells such as cell cycle, apoptosis, growth, and signal transduction. It is one of the most common mechanisms of regulating protein function. Bacteria use two component system consisting of histidine protein

kinases and the associated regulators for phosphorylation of proteins (Stock et al., 1989).

In eukaryotes, phosphorylation is mostly observed at serine, threonine, or tyrosine amino acids with the help of enzymes called serine threonine protein kinases (STPKs). Studies have shown that "eukaryotic like" STPKs are also present in *M. tuberculosis* genome and are involved in several important activities of the cell. For example, PknF was found to be involved in glucose transport and cell division (Deol et al., 2005). The detailed introduction of STPKs of *M. tuberculosis* and *M. marinum* is given in chapter 7 (section 7.1).

1.4.5 Expression of Cpn proteins in mycobacteria

As shown in section 1.4.2 that *cpn60.1* and *cpn10* are always found in a single operon whereas *cpn60.2* is found separately on chromosome. It was observed that despite being in a single operon, the expression levels of *cpn10* and *cpn60.1* were found to be remarkably different in *M. tuberculosis* and *M. smegmatis* (Aravindhan et al., 2009; Rao & Lund, 2010). The spacer region between *cpn10* and *cpn60.1* of *M. tuberculosis* is 94 bps long but there is no evidence of the presence of promoter upstream to Mtb *cpn60.1*. This is further reviewed in chapter 4 sections 4.1 and 4.2. The promoter regions and expression levels of chaperonins of *M. marinum* are not studied yet.

1.4.6 Role in immunomodulation

The chaperonins of mycobacteria are observed to be immunogenic and eliciting the immune response. The 65 kDa protein of *M. bovis* BCG was observed to have high immunogenic activity. It was found to elicit cutaneous reactions in infected guinea pigs (Bruyn et al., 1987). *M. tuberculosis* Cpn10 was observed to have bone resorption activity which was as potent as the strongest osteolytic cytokine IL-1 (Meghji et al., 1997). Both the Cpn60 proteins of *M. tuberculosis* were observed to have high immunoreactive activities and that they not only possess cross reactive epitopes but also genus specific epitopes (Shinnick et al., 1988).

It was also reported that after antibiotic therapy the concentration of antibodies specific to *M. tuberculosis* Cpn60.2 rises in the serum of the host and persist for nearly 18 months post treatment (Kamal, 1998). In autoimmune diseases like systemic

sclerosis and rheumatoid arthritis, an increased level of Cpn60.2 specific antibodies was observed (Eden et al., 1998). Although not required for cell survival, the Cpn60.1 of *M. tuberculosis* was found to be essential for granulomatous inflammation of infected mice and guinea pigs as discussed in section 1.1.2 (Figure 1.5). In summary, the Cpn proteins of *M. tuberculosis* have extensively shown to provoke pro and anti-inflammatory immune responses.

1.5 Aims and objectives

The aim of this project is the better understanding of chaperonins of *M. marinum* which we used as a model organism to study the pathogenesis of *M. tuberculosis*. Following are the different topics investigated in this project.

- Genome comparison analysis and phylogenetic analysis of *M. marinum* 1218R, *M. tuberculosis* and other mycobacterial species was performed for better understanding of the model organism. The multiple sequence alignment of the chaperonin genes of *M. marinum* and *M. tuberculosis* was also performed to find the regions of similarity and diversity.
- To investigate the promoter regions of chaperonins of *M. marinum*, the upstream regions of chaperonin genes were studied both qualitatively and quantitatively to demonstrate promoter activity. For this purpose, different lengths of upstream regions of *cpn* genes of *M. marinum* were cloned in a promoter-less *lacZ* expression plasmid. LacZ assay was performed in *M. marinum* 1218R strain under normal and different stressed conditions to find the presence of promoter activity and the effect of stress conditions such as high temperature and pH changes on the activity of *cpn* promoter.
- The effect of deletion of *cpn60.1* on the expression levels of *cpn10* and *cpn60.2* was studied by analysing the activity of *cpn10* and *cpn60.2* promoters in *M. marinum* Δ*cpn60.1* strain.
- The effect of deletion of *cpn60.1* on the transcriptome profiling of the *M. marinum* was studied using RNA-seq analysis of WT, Δ*cpn60.1,* and complemented strains of *M. marinum*. Gene ontology pathway analysis of the significantly differentially genes in Δ*cpn60.1* compared to the WT was also performed to find the role of differentially expressed genes in the cell. The comparative expression levels of the chaperonin genes of *M. marinum* were also observed by calculating the TPM values of the read counts of the triplicate WT samples.
- The functional analysis of chaperonins of *M. marinum* was performed to investigate their protein folding activity *in vivo*. For this purpose, expression and complementation assays were performed in *E. coli* strains having *groESL* operon under tight control of arabinose operon or chromosomal *groEL* deleted (growth was sustained by displaceable plasmid carrying *groEL*). The expression

optimization and RBS distance optimization was done to get the better expression of the chaperonin genes in *E. coli* where necessary.

- The effect of phosphorylation of different residues and ATPase mutations on the complementation ability of the double phosphomimetic mutant of *cpn60.1* of *M. marinum* was also studied. For this purpose, single and double phosphomimetic mutants of *M. marinum cpn60.1* were created and tested for the chaperone function in *E. coli*. Finally, the mass spectrometry analysis of MmCpn60.1 was also tried to find the post-translational modifications.

2 Materials and methods

2.1 Bacterial strains

Following is the list of bacterial strains used in this study.

SN	Strain	Explanation and Genotype	Resistance marker	Source / Reference
1.	TOP10	*E. coli* strain used for high efficiency cloning and plasmid propagation. Genotype similar to DH10B F⁻*mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K λ⁻*rps*L (Str^R) *end*A1 *nup*G.	Str^R	Lab Stock, (Grant et al., 1990)
2.	JM110	*E. coli* strain deficient of Dam /Dcm methylases used for preparation of plasmids to be digested by methylation sensitive restriction enzymes.	No resistance	Lab Stock
3.	*M. marinum* 1218R	*M. marinum* 1218R strain containing a natural plasmid pMMRN.	No resistance	Lab Stock
4.	*M. marinum* 1218R_V	*M. marinum* 1218R strain (pMMRN natural plasmid) + pSK101 plasmid.	Kan^R	Dr C.M. Santosh Kumar
5.	*M. marinum* 1218R_Δ*cpn60.1*	*M. marinum* 1218R strain (pMMRN natural plasmid) + pSK101 containing Kan^R gene + hygromycin resistance cassette and *sacB* gene replacing the chromosomal *cpn60.1*.	Hyg^R, Kan^R	Dr C.M. Santosh Kumar
6.	M. *marinum* 1218R_Comp	*M. marinum* 1218R_Δ*cpn60.1* strain containing functional *cpn60.1* on pSK101 plasmid.	Hyg^R, Kan^R	Dr C.M. Santosh Kumar
7.	M. *smegmatis* MC^2155	*Mycobacterium smegmatis* laboratory wild-type strain (MC^2155)	No resistance	Lab stock
8.	MGM100	Derivative strain of *E. coli* K-12 MG1655, *groE* promoter of *groESL* operon replaced by *araBAD* promoter (p*BAD*) and *araC* gene. A kanamycin resistance cassette inserted downstream of *groEL.*	Kan^R	Lab stock

9.	AI90	*E. coli* strain in which the chromosomal copy of *groEL* gene (Δ*groEL*) has been replaced by kanamycin resistance cassette. Growth is maintained by *trc* regulated *groESL* on an expression plasmid.	Kan^R, Cm^R	Lab stock
10.	Tab21	Derived from BL21 (DE3) in which T7 RNA polymerase gene is under control of *lac* promotor. In Tab21 *groE-groESL* operon of BL21 has been replaced with p*BAD-groESL-kanR* operon used in MGM100.	Kan^R	Lab stock
11.	*M. marinum* E11	*M. marinum* E11 strain	No resistance	Lab stock

Table 2.1 List of bacterial strains used in this study.

2.2 Plasmids

Following is the table of primers used in this study.

SN	Plasmid	Explanation	Resistance marker	Source / Reference
1.	pUC18	2686 bps high copy number *E. coli* cloning vector. Used for blue-white screening.	Amp^R	Lab stock
2.	pUC18_10_S	pUC18 containing 203 bps of upstream region of *cpn10*	Amp^R	This study
3.	pUC18_10_L	pUC18 containing 313 bps of upstream region of *cpn10*	Amp^R	This study
4.	pUC18_10_XL	pUC18 containing 369 bps of upstream region of *cpn10*	Amp^R	This study
5.	pUC18_60.1_S	pUC18 containing 207 bps of upstream region of *cpn60.1*	Amp^R	This study
6.	pUC18_60.1_L	pUC18 containing 401 bps of upstream region of *cpn60.1*	Amp^R	This study

7.	pUC18_60.2	pUC18 containing 309 bps of upstream region of *cpn60.2*	Amp^R	This study
8.	pSD5B	Promoter-selection *E. coli*–Mycobacterium shuttle vector containing a promoter-less *lacZ* gene	Kan^R	Lab stock, (Jain et al., 1997)
9.	p5B_10_S	pSD5B containing 203 bps of upstream region of *cpn10*	Kan^R	This study
10.	p5B_10_L	pSD5B containing 313 bps of upstream region of *cpn10*	Kan^R	This study
11.	p5B_10_XL	pSD5B containing 369 bps of upstream region of *cpn10*	Kan^R	This study
12.	p5B_60.1_S	pSD5B containing 207 bps of upstream region of *cpn60.1*	Kan^R	This study
13.	p5B_60.1_L	pSD5B containing 401 bps of upstream region of *cpn60.1*	Kan^R	This study
14.	p5B_60.2	pSD5B containing 309 bps of upstream region of *cpn60.2*	Kan^R	This study
15.	pBR-*groES*-Mm*cpn60.1*	pBR322 derivative plasmid containing *E. coli groES* and *M. marinum cpn60.1*. Used as a source of *E. coli groES* and *M. marinum cpn60.1.*	Tet^R, Amp^R	Lab stock
16.	pBR-*groES*-Mm*cpn60.2*	pBR322 derivative plasmid containing *E. coli groES* and *M. marinum cpn60.2*. Used as a source of *E. coli groES* and *M. marinum cpn60.2.*	Tet^R, Amp^R	Lab stock
17.	pTrc99A	4176bp pTrc expression vector driven by lactose/IPTG inducible tightly regulated *trc* promoter.	Amp^R	Lab stock, (Amann et al., 1988)
18.	pTrc-*groESL*	pTrc99A derivative plasmid containing *E. coli groES* and *groEL* genes under control of *trc* promoter.	Amp^R	Lab stock
19.	pTrc-Mtb*cpn10*-Mtb*cpn60.1*	pTrc99A derivative plasmid containing *M. tuberculosis cpn10* and *cpn60.1* genes under control of *trc* promoter.	Amp^R	Lab stock

20.	pTrc-Mtb*cpn10*-Mtb*cpn60.2*	pTrc99A derivative plasmid containing *M. tuberculosis cpn10* and *cpn60.2* genes under control of *trc* promoter.	Amp^R	Lab stock
21.	pTrc-Ms*cpn10*-Ms*cpn60.*1	pTrc99A derivative plasmid containing *M. smegmatis cpn10* and *cpn60.1* genes under control of *trc* promoter.	Amp^R	Lab stock
22.	pTrc-Ms*cpn10*-Ms*cpn60.2*	pTrc99A derivative plasmid containing *M. smegmatis cpn10* and *cpn60.2* genes under control of *trc* promoter.	Amp^R	Lab stock
23.	pTrc-*groES*-Mm*cpn60.1*	pTrc99A derivative plasmid containing *E. coli groES* and *M. marinum cpn60.1* genes under control of *trc* promoter.	Amp^R	This study
24.	pTrc-*groES*-Mm*cpn60.2*	pTrc99A derivative plasmid containing *E. coli groES* and *M. marinum cpn60.2* genes under control of *trc* promoter.	Amp^R	This study
25.	pTrc-Mm*cpn10*-Mm*cpn60.1*	pTrc99A derivative plasmid containing *M. marinum cpn10* and *cpn60.1* genes under control of *trc* promoter.	Amp^R	This study
26.	pTrc-Mm*cpn10*-Mm*cpn60.2*	pTrc99A derivative plasmid containing *M. marinum cpn10* and *cpn60.2* genes under control of *trc* promoter.	Amp^R	This study
27.	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO	pTrc99A derivative plasmid containing sequence optimized *M. marinum cpn10* and *cpn60.1* genes under control of *trc* promoter.	Amp^R	This study
28.	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-6RBS	pTrc99A derivative plasmid containing sequence optimized *M. marinum cpn10* and *cpn60.1* genes under control of *trc* promoter, 6 bps distance between RBS and start codon of Mm*cpn60.1*.	Amp^R	This study
29.	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-8RBS	pTrc99A derivative plasmid containing sequence optimized *M. marinum cpn10* and *cpn60.1* genes under control of *trc* promoter, 8 bps distance between RBS and start codon of Mm*cpn60.1*.	Amp^R	This study

30.	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-10RBS	pTrc99A derivative plasmid containing sequence optimized *M. marinum cpn10* and *cpn60.1* genes under control of *trc* promoter, 10 bps distance between RBS and start codon of Mm*cpn60.1*.	Amp^R	This study
31.	pTrc-Mtb*cpn10*-Mtb*cpn60.1*-DPM	pTrc99A derivative plasmid containing *M. tuberculosis cpn10* and *cpn60.1* genes under control of *trc* promoter, *cpn60.1* gene contains two mutations S75D and S393D (double phosphomimetic mutant).	Amp^R	Dr C.M. Santosh Kumar
32.	pTrc-Mm*cpn10*-Mm*cpn60.*1-CO-S75D	pTrc99A derivative plasmid containing *M. marinum cpn10* and *cpn60.1* sequence optimized genes under control of *trc* promoter. *cpn60.1* gene contains one mutation S75D (single phosphomimetic mutant).	Amp^R	This study
33.	pTrc-Mm*cpn10*-Mm*cpn60.*1-CO- S393D	pTrc99A derivative plasmid containing *M. marinum cpn10* and *cpn60.1* sequence optimized genes under control of *trc* promoter. *cpn60.1* gene contains one mutation S393D (single phosphomimetic mutant).	Amp^R	This study
34.	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-6RBS	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-6RBS plasmid, Mm*cpn60.1* gene contains two mutations S75D and S393D (double phosphomimetic mutant).	Amp^R	This study
35.	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-8RBS	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-8RBS plasmid, Mm*cpn60.1* gene contains two mutations S75D and S393D (double phosphomimetic mutant).	Amp^R	This study
36.	pTrc-Mm*cpn10*-Mm*cpn60.*1-CO-DPM-10RBS	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-10RBS plasmid, Mm*cpn60.1* gene contains two mutations S75D and S393D (double phosphomimetic mutant).	Amp^R	This study
37.	pTrc-Mtb*cpn10*-Mtb*cpn60.1*-CODPM-S415D-8RBS	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-8RBS plasmid with an additional mutation of S415D (Triple phosphomimetic mutant).	Amp^R	This study
38.	pTrc-Mtb*cpn10*-Mtb*cpn60.1*-CO-DPM-S443D-8RBS,	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-8RBS plasmid with an additional mutation of S443D (Triple phosphomimetic mutant).	Amp^R	This study
39.	pTrc-Mm*cpn10*- Mm*cpn60.1*-CO-DPM-T483D-8RBS	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-8RBS plasmid with an additional mutation of S483D (Triple phosphomimetic mutant).	Amp^R	This study

40.	pTrc-Mm*cpn10*-Mm*cpn60.*1-CO-DPM-T481D-8RBS	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-8RBS plasmid with an additional mutation of S481D (Triple phosphomimetic mutant).	Amp^R	This study
41.	pTrc- Mmcpn10-Mmcpn60.1-CO-DPM-T481D+T483D-8RBS.	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-8RBS plasmid with two additional mutation of S481D+S483D (Quadruple phosphomimetic mutant).	Amp^R	This study
42.	pTrc-*groES*-Mm*cpn60.1*-CO	pTrc99A derivative plasmid containing *E. coli groES* and sequence optimized *cpn60.1* genes under control of *trc* promoter.	Amp^R	This study
43.	pTrc-*groES*-Mm*cpn60.1*-CO-DPM	pTrc-*groES*-Mm*cpn60.1*-CO plasmid, Mm*cpn60.1* gene contains two mutations S75D and S393D (double phosphomimetic mutant).	Amp^R	This study
44.	pTrc-*groES*-Mm*cpn60.*1-CO-DPM 8RBS	pTrc-*groES*-Mm*cpn60.1*-CO-8RBS plasmid, Mm*cpn60.1* gene contains two mutations S75D and S393D (double phosphomimetic mutant).	Amp^R	This study
45.	p_CO-D51A	pTrc-*groES*-Mm*cpn60.1*-CO plasmid, Mm*cpn60.1* gene contains one mutation D51A (single ATPase mutant).	Amp^R	This study
46.	p_CO-D396A	pTrc-*groES*-Mm*cpn60.1*-CO plasmid, Mm*cpn60.1* gene contains one mutation D396A (single ATPase mutant).	Amp^R	This study
47.	p_CO-D51A+D396A	pTrc-*groES*-Mm*cpn60.1*-CO plasmid, Mm*cpn60.1* gene contains two mutations D51A + D396A (double ATPase mutant).	Amp^R	This study
48.	p_CO+D51A+D86A+D396A	pTrc-*groES*-Mm*cpn60.1*-CO plasmid, Mm*cpn60.1* gene contains three mutations D51A + D86A + D396A (triple ATPase mutant).	Amp^R	This study
49.	p_CO-DPM-D86A	pTrc-*groES*-Mm*cpn60.1*-CO-DPM plasmid, Mm*cpn60.1* gene contains one mutation D86A (single ATPase mutant).	Amp^R	This study
50.	p_DO-DPM-D51A	pTrc-*groES*-Mm*cpn60.1*-CO-DPM plasmid, Mm*cpn60.1* gene contains one mutation D51A (single ATPase mutant).	Amp^R	This study

51.	p_CO-CPM-8- D51A	pTrc-*groES*-Mm*cpn60.*1-CO-DPM 8RBS plasmid, Mm*cpn60.1* gene contains one mutation D51A (single ATPase mutant).	Amp^R	This study
52.	p_CO-DPM-8-D86A	pTrc-*groES*-Mm*cpn60.*1-CO-DPM 8RBS plasmid, Mm*cpn60.1* gene contains one mutation D86A (single ATPase mutant).	Amp^R	This study
53.	p_CO-DPM-8-D51A+D86A	pTrc-*groES*-Mm*cpn60.*1-CO-DPM 8RBS plasmid, Mm*cpn60.1* gene contains two mutations D51A + D86A (double ATPase mutant).	Amp^R	This study
54.	p_CO-DPM-8-F	pTrc-*groES*-Mm*cpn60.*1-CO-DPM 8RBS plasmid, Mm*cpn60.1* gene contains a frameshift mutation.	Amp^R	This study
55.	pTrc-Mtb*cpn10*-Mtb*cpn60.1*_MmHindIII_DPM	pTrc-Mtb*cpn10*-Mtb*cpn60.1*-DPM having 531 bps of Mtb*cpn60.1*-DPM from C-terminal swapped with Mm*cpn60.1*-CO- DPM.	Amp^R	This study
56.	pTrc-Mm*cpn10*-Mm*cpn60.1*_MtbHindIII_DPM	pTrc-Mm*cpn10*-Mm*cpn60.1*-CO_DPM_8RBS having 531 bps of Mm*cpn60.1*-CO-DPM from C-terminal swapped with Mtb*cpn60.1*-DPM.	Amp^R	This study
57.	pTrc-*groES*-SpeI	pTrc99A derivative plasmid containing *E. coli groES* under regulation of *trc* promoter.	Amp^R	Lab Stock

Table 2.2 Table of plasmids used in this study.

2.3 Primers

SN	Name of the Primer/oligo	Use	Sequence 5′ to 3′	Deletion /Insertion	No of bases
1.	cpn10_S_F	Forward primer for amplification of 203 bps upstream to 1218R *cpn10.*	TAAGCATCTAGAGCAGTCGGACGACCCCTGCG	XbaI	32
2.	cpn10_L_F	Forward primer for amplification of 313 bps upstream to 1218R *cpn10.*	TAAGCATCTAGATCCCGGTGCGGCCGCGGC	XbaI	30
3.	cpn10_XL_F	Forward primer for amplification of 369 bps upstream to 1218R *cpn10.*	TAAGCATCTAGACGCTGGATGTGCCCAGCGAC	XbaI	32
4.	cpn10_S/L/XL /_R	Reverse primer for amplification of 203, 313, and 369 bps upstream to 1218R *cpn10*.	TCCGTAGCATGCGATTGGAGCCCTCCACTAATTC	SphI	34
5.	cpn60.1_S_F	Forward primer for amplification of 207bps upstream to 1218R *cpn60.1.*	TAAGCATCTAGAGGCAGAGGGTGACACCGTC	XbaI	31
6.	cpn60.1_L_F	Forward primer for amplification of 401 bps upstream to 1218R *cpn60.1.*	TAAGCATCTAGACCAATCGTGGCGAAGGTGAAC	XbaI	33
7.	cpn60.1_R	Reverse primer for amplification of 207 and 401 bps upstream to 1218R *cpn60.1*.	TCCGTAGCATGCCAGGCTCCTTATGCCCAGC	SphI	31
8.	cpn60.2_F	Forward primer for amplification of 309 bps upstream to 1218R *cpn60.2.*	TAAGCATCTAGACACCCCAGTGTGGCCCAACG	XbaI	32
9.	cpn60.2_R	Reverse primer for amplification of 309 bps upstream to 1218R *cpn60.2*.	TCCGTAGCATGCTGCGAAGTGATTCCTCCGG	SphI	31

Table 2.3 Primers used for amplification of upstream regions of *M. marinum* 1218R cpn10, cpn60.1 and cpn60.2 for cloning into pSD5B.

SN	Name of the Primer/oligo	Use	Sequence 5′ to 3′	No of bases
10.	SphI_F	Forward oligos	CGTTAGCGGACTAGCTGGGCATAAGGAGCCTGCATG	36
11.	SphI_R	Reverse oligos	CAGGCTCCTTATGCCCAGCTAGTCCGCTAACGCATG	

Table 2.4 Oligos for insertion of missing 31 bps in p5B_60.1_S and p5B_60.1_L plasmid constructs (Figure 4.23).

SN	Name of the Primer/oligo	Use	Sequence 5′ to 3′	Deletion /Insertion	No of bases
12.	cpn10_F	Forward primer for amplification of Mm*cpn10* gene.	AGGGCTCCCATGGTGGCGAAGGTGAACATCAAGC	NcoI	34
13.	cpn10_R	Reverse primer for amplification of Mm*cpn10* gene.	CGGGGATAGATCTGGCGGTACACGGGGTTACTTG	BglII	

Table 2.5 Primers used for amplification of cpn10 gene from *M. marinum* genomic DNA to clone into pTrc99A between NcoI and BglII sites.

SN	Name of the Primer/oligo	Use	Sequence 5′ to 3′	Deletion /Insertion	No of bases
14.	60.1_NcoI_F	Forward primer used for SDM.	GACGAAACTGCGCGCCGCGCAATGGAGGCGGGCGTGGAC	GCC to GCA (Alanine)	39
15.	60.1_NcoI_R	Reverse primer used for SDM.	GTCCACGCCCGCCTCCATTGCGCGGCGCGCAGTTTCGTC		

Table 2.6 Primers used for removal of intrinsic NcoI site from cpn60.1 gene of *M. marinum* in pTrc-groES-Mmcpn60.1.

SN	Name of the Primer/oligo	Use	Sequence 5′ to 3′	Deletion /Insertion	No of bases
16.	6RBS_F	Forward primer used for SDM to inset AT between RBS and start codon of Mm*cpn60.1*.	CTCTGGGCATAAGGAGATCCTGATGAGCAAGCTGATCGAG	AT	40
17.	6RBS_R	Reverse primer used for SDM to insert AT.	CTCGATCAGCTTGCTCATCAGGATCTCCTTATGCCCAGAG		
18.	8RBS_F	Forward primer used for SDM to inset ATAT between RBS and start codon of Mm*cpn60.1*.	CTCTGGGCATAAGGAGATATCCTGATGAGCAAGCTGATCGAG	ATAT	42
19.	8RBS_R	Reverse primer used for SDM to inset ATAT.	CTCGATCAGCTTGCTCATCAGGATATCTCCTTATGCCCAGAG		
20.	10RBS_F	Forward primer used for SDM to inset between RBS and start codon of Mm*cpn60.1* ATATAC.	CTCTGGGCATAAGGAGATATACCCTGATGAGCAAGCTGATCGAG	ATATAC	44
21.	10RBS_R	Reverse primer used for SDM to inset ATATAC.	CTCGATCAGCTTGCTCATCAGGGTATATCTCCTTATGCCCAGAG		

Table 2.7 Primers used for site directed mutagenesis to change the distance of RBS in pTrc- Mmcpn10-Mmcpn60.1 CO (Figure 6.15).

SN	Name of the Primer/oligo	Use	Sequence 5′ to 3′	Deletion /Insertion	No of bases
22	S75D_F	Forward and reverse primers used for SDM to convert AGC (Serine) to GAC (Aspartic acid) at position 75 of MmCpn60.1.	GGTGCGCAGCTGGTTAAGGACGTGGCGACCAAAACCAAC	AGC to GAC	39
23	S75D_R		GTTGGTTTTGGTCGCCACGTCCTTAACCAGCTGCGCACC		
24	S393D_F	Forward and reverse primers used for SDM to convert AGC (Serine) to GAT (Aspartic acid) at position 393 of MmCpn60.1.	CTGAAGAAACGTAAAGAGGATGTGGAAGATGCGGTGGCGGCG	AGC to GAT	42
25	S393D_R		CGCCGCCACCGCATCTTCCACATCCTCTTTACGTTTCTTCAG		
26	S415D_F	Forward and reverse primers used for SDM to convert AGC (Serine) to GAC (Aspartic acid) at position 415 of MmCpn60.1.	TATTGTTGCGGGCGGTGGCGACGCGCTGATTCAGGCGCGTAA	AGC to GAC	42
27	S415D_R		TTACGCGCCTGAATCAGCGCGTCGCCACCGCCCGCAACAATA		
28	S443D_F	Forward and reverse primers used for SDM to convert AGC (Serine) to GAC (Aspartic acid) at position 443 of MmCpn60.1.	TGGGTGTTGATGTGTTCGACGAGGCGCTGGCGGCGC	AGC to GAC	36
29	S443D_R		GCGCCGCCAGCGCCTCGTCGAACACATCAACACCCA		
30	T483D_F	Forward and reverse primers used for SDM to convert ACC (Threonine) to GAC (Aspartic acid) at position 483 of MmCpn60.1.	TGAACGCGGCGACCCTGGACTATGGCGACCTGGCGGC	ACC to GAC	37
31	T483D_R		GCCGCCAGGTCGCCATAGTCCAGGGTCGCCGCGTTCA		
32	T481D_F	Forward and reverse primers used for SDM to convert ACC (Threonine) to GAT (Aspartic acid) at position 481 of MmCpn60.1.	TCTGAACGCGGCGGATCTGACCTATGGCGACCTGG	ACC to GAT	35
33	T481D_R		CCAGGTCGCCATAGGTCAGATCCGCCGCGTTCAGA		

34	T481+482D_F	Forward and reverse primers used for SDM to convert ACC (Threonine) to GAT and GAC (Aspartic acid) at positions 481 and 482 of MmCpn60.1.	TGAACGCGGCGGATCTGGACTATGGCGACCTGGCGGC	ACC to GAT and GAC	37
35	T481+482D_R		GCCGCCAGGTCGCCATAGTCCAGATCCGCCGCGTTCA		

Table 2.8 Primers used for construction of phosphomimetic mutations.

SN	Name of the Primer/oligo	Use	Sequence 5′ to 3′	Deletion /Insertion	No of bases
36	D51A_F	Forward and reverse primers used for SDM to convert GAT (Aspartic acid) to GCG (Alanine) at position 51 of MmCpn60.1.	TGGCCCGACCGTTACCAACGCGGGCGTGACCGTTGCGCGT	GAT to GCG	40
37	D51A_R		ACGCGCAACGGTCACGCCCGCGTTGGTAACGGTCGGGCCA		
38	D86A_F	Forward and reverse primers used for SDM to convert GAT (Aspartic acid) to GCG (Alanine) at position 86 of MmCpn60.1.	AACCAACGATGTGGCGGGTGCGGGTACCACCACCGCGACCGT	GAT to GCG	42
39	D86A_R		ACGGTCGCGGTGGTGGTACCCGCACCCGCCACATCGTTGGTT		
40	D396A_F	Forward and reverse primers used for SDM to convert GAT (Aspartic acid) to GCG (Alanine) at position 396 of MmCpn60.1.	ACGTAAAGAGAGCGTGGAAGCGGCGGTGGCGGCGGCGAAA	GAT to GCG	40
41	D396A_R		TTTCGCCGCCGCCACCGCCGCTTCCACGCTCTCTTTACGT		

Table 2.9 Primers used for construction of ATPase mutations.

SN	Name of the Primer/oligo	Use	Sequence 5′ to 3′	No of bases
42	M13_F	Forward primer for sequencing of fragments cloned in pUC18.	CCCAGTCACGACGTTGTAAAACG	23
43	M13_R	Reverse primer for sequencing of fragments cloned in pUC18.	AGCGGATAACAATTTCACACAGG	23
44	cpn60.1_S/L_R	Reverse primer for screening of successfully cloned pUC18 derivatives of the two *cpn60.1* upstream fragments.	GCATGCCGCCCCGGAAATCAC	21
45	pSD5B_R	Reverse primer for sequencing of fragments cloned in pSD5B.	GCAAGGCGATTAAGTTGG	18
46	pTrc99A_F	Forward primer for sequencing of fragments cloned in pTrc99A.	GCACTCCCGTTCTGGATAAT	20
47	pTRc99A_R	Reverse primer for sequencing of fragments cloned in pTrc99A.	GCGTTTCACTTCTGAGTTCG	20
48	cpn60.1_int_R	Reverse primers within Mm*cpn60.1*.	GATGATGGTGTCGTCCTTG	19
49	cpn60.2_int1_R	Reverse primer in Mm*cpn60.2*	GTAGTCGGAGTCGCTGTTCT	20
50	cpn60.2_int2_R	Reverse primer in Mm*cpn60.1* close to 5' end of gene	GTCGTACGCAATTGTCTTGG	20

Table 2.10 Primers used for sequencing and screening.

SN	Name of the Primer/oligo	Use	Sequence 5′ to 3′	Insertion	No of bases
51	BglII_F	Forward primer used for SDM.	GTACGCGTAGGCGTCTGAA**AGATCT**TTAATCCGCGCACGACAC	AGATCT (BglII)	43
52	BglII_R	Reverse primer used for SDM.	GTGTCGTGCGCGGATTAAAGATCTTTCAGACGCCTACGCGTAC		

Table 2.11 Primers used SDM to insert BglII site (chapter 6, figure 6.18).

2.4 General microbiology techniques

2.4.1 Growth media

2.4.1.1 LB Broth and LB agar

To prepare Luria-Bertani (LB) broth media, (10 g/L) 1% w/v tryptone, (5 g/L) 0.5% w/v yeast extract, and (10 g/L) 1% NaCl was dissolved in distilled water (pH was adjusted to 7.0 by using NaOH solution) and sterilized by autoclaving at 15 psi and 121°C for 15 minutes. To prepare LB agar, 15 g per litre (1.5%) agar was added to the LB broth before autoclaving.

2.4.1.2 Difco 7H9 and 7H10

4.7 g of Difco™ Middlebrook 7H9 broth from BD Biosciences was added to 900 ml of distilled water with addition of 0.05% tween-80 or 2 ml glycerol and sterilized by autoclaving. For solid media, 19 g powder of Difco™ Middlebrook 7H10 agar from BD Biosciences was added to 900 ml of distilled water containing 5ml of glycerol, mixed thoroughly and sterilized by autoclaving. 11.11% Middlebrook OADC (100 ml per 900 ml) was added to 7H9 broth and 7H10 agar aseptically before use (or when cooled to 50-55°C).

2.4.1.3 Super Optimal Broth (SOB)

SOB media was prepared by dissolving 2% tryptone, 0.5% yeast extract, 10mM NaCl, 2.5 mM KCL, 10mM $MgCl_2$, and 10 mM $MgSO_4$ in distilled water. The solution was sterilized by autoclaving.

2.4.1.4 SOC media

SOC is identical to SOB with one additional supplement i.e., glucose. For SOC media preparation, SOB media was cooled to 60°C and filter sterilized glucose solution was added (20 mM final glucose concentration).

2.4.2 Media supplements

The stock and working concentrations of all the supplements are given in Table 2.12. All the stock solutions were prepared in Ultrapure water (UPW) water and filtered using 0.2 mm diameter sterile syringe filter unless stated otherwise. All the media and reagents used were sourced from Sigma-Aldrich, UK.

Sr No.	Supplement	Stock concentration	Working concentration
1.	L-arabinose	20% (w/v)	0.2% (w/v)
2.	D-Glucose	20% (w/v)	0.2% (w/v)
3.	D-Lactose	20% (w/v)	0.2% (w/v)
4.	Sucrose	50% (w/v)	2% (w/v)
5.	IPTG	1 M	0.1 mM
6.	X-Gal (Solvent: DMF)	20 mg/ml	200 µg/ml
7.	Ampicillin (Amp)	100 mg/ml	100 µg/ml
8.	Kanamycin (Kan)	50 mg/ml	50 µg/ml
9.	Tetracycline (Tet)	10 mg/ml	10 µg/ml
10.	Hygromycin (Hyg)	100 mg/ml	25 µg/ml
11.	Chloramphenicol (Cam) (Solvent: Ethanol)	50 mg/ml	50 µg/ml

Table 2.12 Stock and working concentrations of supplements used in this study.

2.4.3 Growth conditions

2.4.3.1 *E. coli* growth conditions

All the *E. coli* strains were grown in LB media. For primary cultures, a single colony was inoculated into 5 ml of LB broth in a sterile culture tube and incubated overnight at 30°C or 37°C in a 180 rpm shaking incubator. For secondary growth culture, 0.1% of the secondary growth volume was used from the overnight culture for inoculation. Growth was monitored by measuring the optical density of the cultures at a wavelength

of 600 nm (OD_{600}). LB agar plates were used for *E. coli* solid media requirements and incubated at various temperatures. Suitable supplements were added in all the media as required.

2.4.3.2 Mycobacteria growth conditions

Mycobacterial strains were grown using 7H9 or 7H10 media supplemented with OADC and suitable additive. *M. marinum* liquid and solid media cultures were grown at 30°C for 7 days (or up to OD_{600} between 0.8 to 1.0 – mid log phase) unless stated otherwise. *M. smegmatis* cultures were grown at 37°C for 3 days.

2.4.4 Glycerol stocks

To store the bacterial cultures, 500 ml of overnight culture was mixed gently with 500 ml of 50% sterile glycerol (1:1) in a 2 ml cryovial and frozen at -70°C.

2.4.5 Genomic DNA extraction

Following method was used to extract genomic DNA of *M. marinum* 1218R and E11 strains.

Reagents required:

1. TE buffer (10mM Tris HCl (pH 8.0), 1mM EDTA): 1 ml of 1M Tris-HCl and 200 µl of 0.5M EDTA was added to 98.8 ml of distilled water and sterilized by autoclaving.
2. Lysozyme powder
3. RNase A
4. P1 buffer (P1 buffer of QIAGEN Miniprep Kit was used)
5. 10% SDS: 20 g of Sodium dodecyl sulfate was added to distilled water 200 ml (final vol.) of and filter sterilized using 0.2 mm filter.
6. Proteinase K (10mg/ml): 500 µl of 1M Tris HCl and 100 µl of 200mM $Ca(CH_3COO)_2$ were added to 9.4 ml sterilized water to make 10 ml solution. 0.1 g of Proteinase K powder was dissolved in the solution, filter sterilized and saved at -20**°C** in aliquots.
7. 5M NaCl: 146.1 g NaCl was dissolved in 500 ml distilled water (final Vol.) and sterilized by autoclaving.

8. Phenol-Chloroform-isoamyl alcohol (25:24:1)
9. Chloroform-isoamyl alcohol (24:1)
10. Isopropanol
11. 70% ethanol

Method:

20 ml secondary growth culture of *M. marinum* 1218R and E11 at mid log phase (OD_{600} 0.8-1.0) was pelleted down by centrifugation at 5000 rpm for 10 minutes, the supernatant was removed carefully, and the pellet was resuspended in 3 ml of TE buffer (pH 8.0). A spatula tip full of lysozyme (powder) was added followed by the addition of 30 µl RNaseA (10 µg/µL) and 250 µl buffer P1. The sample was then placed overnight in a shaking incubator at 37°C. Following day, 420 µl of 10% SDS and 60 µl of proteinase K was added, mixed gently, and incubated at 65°C in the dryer for 2 hours. The sample was taken out of the dryer, 600 µl of 5M NaCl was added and incubated for 30 minutes at room temperature. 4.5 ml Phenol-Chloroform-isoamyl alcohol (25:24:1) was added to the sample and mixed well for 15 minutes. The sample was then centrifuged at 12000 rpm for 20 minutes. The upper aqueous layer was transferred to a new 25 ml falcon tube and 4.5 ml of Chloroform-isoamyl alcohol (24:1) was added. The sample was again centrifuged at 12000 rpm for 20 minutes followed by the transfer of the upper phase to a new 25 ml falcon tube containing 3 ml isopropanol. The tube was inverted gently, a cloud (threads) of genomic DNA was seen in the tube. The DNA was fished into 1.5ml Eppendorf tube (labelled) containing 500 µL of 70% ethanol. The DNA was washed 2 times by centrifugation for 10-15 minutes in 500 µL 70% ethanol. The pellet of genomic DNA was dried at 37°C and resuspended in 100 100 µL TE (pH 8.0).

2.4.6 Agarose gel electrophoresis

Reagents required:

1. Agarose
2. Tris-acetate-EDTA (TAE) buffer: To make 1 L of 50X stock of TAE, 242 g of Tris base, 100 ml of 0.5M EDTA pH 8.0 (or 18.61 g of disodium EDTA) and 57.1 ml of glacial

acetic acid was added to approximately 600 ml of distilled water (DW), mixed thoroughly and final volume was adjusted to 1 L by the addition of DW.

3. Midori green advanced DNA stain
4. DNA gel loading dye 6X (New England BioLabs)
5. HyperLadder™ 1Kb (BioLine)

Method:

DNA (genomic, plasmid, restriction digests, PCR samples etc) was resolved on agarose gels. 1% or 0.8% agarose gel was prepared using 1X TAE with addition of Midori green DNA stain (5 ml / 100 ml). DNA loading dye was added to the DNA sample (5 ml per 20 ml) before loading into the wells. The gel was run in a tank containing 1X TAE buffer at 100 V for about 1 – 1.5 hours.

2.4.7 Plasmid purification

Two methods were used for purification of plasmids.

2.4.7.1 Alkaline lysis method

Reagents required:

1. GTE buffer: 50mM glucose, 25mM TRIS pH 8.0, 10mM EDTA (autoclaved)
2. 1%SDS/0.2M NaOH: (Made fresh for every use from stock solutions of 10%SDS and 10M NaOH)
3. 3MNa/5M Acetate: 3M CH_3COONa, adjusted pH 5.0 with glacial acetic acid
4. TE: 10mM TRIS pH8.0, 1mM EDTA (autoclaved)
5. Isopropanol
6. 70% ethanol
7. 100% ethanol

Method:

10 ml overnight culture was centrifuged at 7000 rpm for 10 minutes to harvest the cells. The supernatant was poured off and the remaining drops of media were removed by pipette. The pellet was resuspended in 200 µl of GTE buffer. For cell lysis and denaturation of chromosomal DNA, 400 µl of 1% SDS/0.2M NaOH was added to the

resuspended pellet followed by addition of 300 µl of 3MNa/5M Acetate. A white insoluble precipitate containing chromosomal DNA, proteins, lipids, and other cell debris was formed. The reaction tube was incubated for 5 minutes at room temperature followed by centrifugation at 13000 rpm for 10 minutes. The resulted supernatant, containing the plasmid DNA, was pipetted out and transferred to an Eppendorf tube containing 600 µl isopropanol. The tube was inverted several times to precipitate the nucleic acid. The tube was centrifuged at 13000 rpm for 10 minutes to collect the small pellet of nucleic acid. The pellet was washed twice with 70% and 100% ethanol. Finally, pellet was dried by placing the tube open at 37^{o}C and resuspended in 50 µl of TE.

2.4.7.2 Plasmid purification using QIAGEN Plasmid kits

For small scale purification of plasmids QIAGEN Plasmid Mini kit was used whereas for large scale plasmid preparations QIAGEN Plasmid Midi kit was used according to the guidelines provided by the manufacturers. For mini prep, 10 ml of overnight culture was spun down to harvest the cells. The pellet was resuspended in 250 µl of buffer P1, mixed well by vortexing and transferred to a sterile microcentrifuge tube. For cell lysis 250 µl of buffer P2 was added to the tube and mixed gently by inverting tube. Immediately after mixing 350 µl of N3 buffer was added to the tube. White insoluble precipitate containing cell debris was pelleted down by centrifugation at 13K rpm for 15 minutes. The supernatant containing plasmid DNA was added to the QIAPrep column. The column was spun down at 7000 rpm for 20 seconds (plasmid DNA was now bound to the column), followed by washing up of the column twice with 750 µl PE buffer. The plasmid DNA was eluted in 50 µl of elution buffer (EB).

2.4.8 Preparation of chemically competent cells

Competent cells of *E. coli* strains were prepared using two methods. The first $CaCl_2$ method was used for preparation of chemically competent cells having medium to high efficiency. These cells were used for propagation of plasmids. The second Inoue method was used for preparation of ultra-competent cells which were used for transformation of ligated DNA.

2.4.8.1 Calcium chloride method

Reagents required:

1. 1M $CaCl_2$ (stock solution): 11.1 g $CaCl_2$ (Mol. Mass 110.98 g) was dissolved in final volume of 100 ml ddH_2O and sterilized by passing through 0.2 mm filter and stored in refrigerator.
2. 50% glycerol (stock Solution): 50% glycerol solution was autoclaved for sterilization.
3. 0.1 M $CaCl_2$ (working Solution): 50 ml of 1M $CaCl_2$ stock was mixed in ddH_2O to make the final volume 500 ml.
4. 0.1 M $CaCl_2$-15% glycerol (working solution): 6 ml of 1M $CaCl_2$ and 18 ml of 50% sterile glycerol was mixed in ddH_2O to make final volume 60 ml.

All solutions were stored in refrigerator, just before use the working solutions were kept on ice.

Method:

A 1 L flask containing 250 ml of liquid media was inoculated with 250 µl of overnight culture, the cells were grown in a 180 rpm shaking incubator at 30°C temperature till the OD_{600} reached to 0.4 – 0.5. The culture was immediately transferred to 50 ml falcons and incubated on ice for 10 minutes. All the subsequent steps were done on ice or at 4°C. Cells were harvested by spinning the falcons at 5000 rpm in a 4°C centrifuge for 15 minutes. The supernatant was discarded, and the pellet was resuspended in 50 ml of pre-chilled 0.1M $CaCl_2$. The cells were incubated on ice for 30 minutes then centrifuged again for 15 minutes. The supernatant was removed carefully. The cell pellet was resuspended in 5 ml of 0.1 M $CaCl_2$-15% glycerol solution. 100 µl aliquots of the cell suspension were made in in prechilled microcentrifuge tubes, snap-frozen in liquid nitrogen and store at -70°C.

2.4.8.2 Inoue competent cells preparation method

Reagents required:

1. 0.5 M PIPES (pH 6.7): 15.1 g of PIPES was dissolved in 80 ml of ddH_2O, pH was adjusted to 6.7 with KOH and enough water was added to make the final volume to 100 ml. The solution was sterilized using 0.45 µm syringe filter. Divided in 20 ml aliquots and stored at -20°C.
2. Inoue transformation buffer: 1 L Inoue transformation buffer was prepared by dissolving the reagents given in Table 2.13 in 800 ml ddH_2O, 20 ml of 0.5 M PIPES was added to the solution and final volume was adjusted to 1 L with ddH_2O. The buffer was sterilized by filtration using 0.45 µm Nalgene™ vacuum filter.
3. Super Optimal Broth (SOB)
4. DMSO (dimethyl sulfoxide)

Reagent	Amount used for 1 L solution	Final concentration
Manganese (II) chloride tetrahydrate	10.88 g	55 mM
Calcium chloride dihydrate	2.20 g	15 mM
Potassium chloride	18.65 g	250 mM
PIPES	20 ml	10 mM

Table 2.13 Reagents for Inoue transformation buffer.

Method:

To prepare the primary culture, from a fresh *E. coli* strain (TOP10) culture plate, a single colony was inoculated into 25 ml of SOB medium in a sterile flask and incubated in 250 – 300 rpm shaker for 6-8 hours at 37°C. For secondary growth culture, two 1 L flasks containing 250 ml of SOB were inoculated with 6 ml (1st flask) and 2 ml (2nd flask) of the primary culture. Both flasks were incubated overnight at 18°C in 180 rpm shaking incubator. Following morning, the OD_{600} of the cultures was monitored. When the OD_{600} of one of the cultures was reached to 0.55, the culture was immediately transferred to 50 ml falcon tubes and incubated on ice for 10 minutes. The 2nd culture was discarded. The cells were harvested by centrifugation at 4000 rpm for 10 minutes

at 4°C. The medium was poured off carefully. The remaining drops of the medium was removed by pipetting. The cells were resuspended in 50 ml of ice-cold Inoue buffer and harvested by centrifugation at 4000 rpm for 10 minutes at 4°C. The buffer was removed, and the cells were resuspended gently in 20 ml of ice-cold Inoue buffer. 1.5 ml of DMSO was added as a cryoprotectant. Mixed carefully by pipetting while keeping the cells on ice all the time. After 10 minutes incubation on ice, the cell suspension was aliquoted in pre-chilled microcentrifuge tubes, snap frozen in liquid nitrogen and stored at -70°C.

2.4.9 Transformation

100 µl aliquot of competent cells tube was thawed on ice for 20 minutes. Up to 150 ng of the transforming DNA (plasmid or ligated DNA) was added to the competent cells and mixed gently by pipetting and kept on ice for 30 minutes. The cells were heat shocked for 90 seconds by transferring the tube to a foam rack placed on a pre-heated 42°C water bath. The tube was rapidly transferred to the ice and incubated for 10 minutes. 1 ml of pre-warmed SOC media (at 37°C) was added to the cells. The cells were incubated at optimum temperature for 60 minutes in 180 rpm shaking incubator. The transformed competent cells were plated on LB agar plates containing appropriate antibiotics and additives (if required). The plates were incubated overnight at optimum growth temperature.

2.4.10 Preparation of electrocompetent cells

Reagent required:

10% glycerol – sterilized by autoclaving.

Method:

To introduce the DNA (e.g., plasmid) into the mycobacteria, electrocompetent cells were prepared. 300 ml of secondary growth culture was grown at optimum temperature till the OD_{600} reached to 0.8-1.0. The culture was transferred to 50 ml falcon tubes and incubated on ice for 1.5 hours. The cells were harvested by spinning the Falcon tubes in 4°C centrifuge at 8000 rpm for 15 minutes. The supernatant was discarded followed by the resuspension of the pellet in 50 ml ice cold 10% glycerol.

The cells were harvested again by centrifugation. The pellet was washed second time with 30 ml and third time with 15 ml of ice cold 10% glycerol. After the third wash, the pellet was resuspended in 3 ml of 10% glycerol, aliquoted in 1.5 ml Eppendorf tubes and stored at -70°C.

2.4.10.1 Electroporation

Electrocompetent cells of mycobacteria were thawed on ice. 100 µl of electrocompetent cells were transferred to a pre-chilled 1.5 ml microcentrifuge tube. 2.5 µl (250 – 500 ng) of salt free plasmid DNA was transferred to the cells and mixed by pipetting. The voltage of Eporator (electroporation apparatus) was set at 2500 V. The mixture of cells and DNA was transferred to pre-chilled 0.2 cm/2 mm diameter electroporation cuvette. The cuvette was tapped to bring the suspension to the bottom, placed in the Eporator (electroporation apparatus) and immediately the electric pulse was given once. The cuvette was removed from the chamber, 1 ml of liquid growth media (7H9 + OADC) was added to the cuvette, the cells were transferred to sterile universal container and incubated at optimum temperature for 2 – 3 hours. The electroporated cells were plated on 7H10 agar plates containing appropriate antibiotics. The time constant for most of the pulses was between 4.9 to 5.7 milliseconds.

2.4.11 RNA extraction

Total RNA was extracted from *M. marinum* 1218R (WT), Δ*cpn60.1* and complementation strains using the QIAGEN RNeasy kit. 10 ml of the secondary growth cultures grown at 30°C with optical density 0.8-1.0 were mixed with double volume of RNAprotect Bacteria reagent (QIAGEN). The cells were then collected by centrifugation at 6000 rpm for 5 minutes. The pellet was resuspended in 700 µl of RLT lysis buffer (Qiagen). The mixture was transferred to Lysing Matrix-E, lysed in Fast-Prep MP Biomedical instrument using protocol for *M. tuberculosis* and centrifuged at 13,000 rpm for 15 minutes. The supernatant was transferred to a microcentrifuge tube (RNase free) containing 470 µl of ethanol. Applied the mixture to RNeasy Mini spin column. Total RNA bound to the column was washed once with 700 µl of RW1 buffer and twice with 500 µl of RPE buffer to remove the contaminants. Finally, the column was

transferred to a fresh tube, RNA was eluted in 100 µl of RNase free water and stored at -70°C.

2.4.11.1 RNA quantification and RIN number calculation

The quantification of RNA samples was done using UV spectrophotometer nanodrop set at RNA-40, λ_{230} nm. The concentration of all the samples was between 150 – 2000 ng/µl. For RIN integrity check, 2 µl of each of the sample was transferred to 0.5 ml microcentrifuge tube and the rest of the samples was stored at -70°C. RNA integrity was checked using Agilent RNA Screen Tape assay performed on Agilent 2200 TapeStation system. Prokaryotic RNA analysis was chosen from the programs and High Sensitivity RNA ScreenTape^{R} was used. The concentration of the RNA samples for Screen Tape was adjusted between 1 – 10 ng/µl by diluting with RNase free water. 2 µl of each RNA sample was mixed with 1 µl of high sensitivity RNA sample buffer, vortexed for 1 minute, spun down briefly and heated at 72°C for 3 minutes. The samples were then placed on ice for 2 minutes and centrifuged briefly to collect the samples in the base of the microcentrifuge tubes. The samples were loaded into the 2200 TapeStation. RNA screen tape was placed in the 2200 TapeStation and the software was run according to the manufacturer's instructions.

2.5 Qualitative and quantitative Assays

2.5.1 X-Gal Qualitative assay/ Blue white screening

Reagents required:

X-Gal stock solution (20 mg/ml): 200 mg of X-Gal (molar mass 408. 63 g/mol) was dissolved in 10 ml DMF, vortexed to dissolve completely, aliquoted in 1.5 ml microcentrifuge tubes, wrapped in Al-foil and stored at -20°C.

Method:

For X-Gal qualitative assay, in case of mycobacterial strains, 7H10 square agar plates were prepared by adding 5.55 ml OADC in 50 ml of 7H10 agar cooled to 50°C along with 500 ml X-Gal solution and appropriate antibiotics. The plates were kept covered with Al foil. For blue-white screening, LB agar plates containing X-Gal solution at a final concentration of 20 mg/ml were prepared.

2.5.2 β-galactosidase activity assay (LacZ assay)

Reagents required:

1. Disodium hydrogen phosphate Na_2HPO_4 0.5 M: 0.5 M Na_2HPO_4 (MW 141.96 g/mol) was prepared by dissolving 3.549 g of the salt in ddH_2O to make the final volume 50 ml.
2. Sodium dihydrogen phosphate NaH_2PO_4 0.5 M: 3.90 g of $NaH_2PO_4.2H_2O$ (MW 156.01 g/mol) was dissolved in ddH_2O to make the final volume 50 ml.
3. Potassium chloride KCl 1M: 3.72 g of KCl (MW 74.55 g/mol) was dissolved in ddH_2O to make the final volume 50 ml.
4. Magnesium sulphate $MgSO_4$ 1M: 6.018g of $MgSO_4$ was dissolved in in ddH_2O to make the final volume 50 ml.
5. Sodium carbonate Na_2CO_3 1M: 5. 29 g of Na_2CO_3 (MW 105.99 g/mol) was dissolved in ddH_2O to make the final volume 50 ml.

   All the above solutions were filter sterilized using 0.2 mm syringe filter.
6. Chloroform
7. 0.1% SDS: 0.05 g SDS powder dissolved in 50 ml ddH_2O.
8. Z-buffer: Z-buffer was prepared according to the Table 2.14.
9. β-mercaptoethanol
10. ONPG (4 mg/ml): 0.04 g of ONPG was dissolved in 10 ml of lacZ buffer (after addition of β-mercaptoethanol).

Reagent	For 50 ml prep	Final concentration
Na_2HPO_4 (0.5M)	6 mL	0.06 M
NaH_2PO_4 (0.5M)	4 mL	0.04 M
KCl (1M)	500 µL	0.01 M
$MgSO_4$ (1M)	50 µL	0.001 M
Pure Water	To a final vol. 50 ml	

Table 2.14 Reagents used for the preparation of lacZ buffer.

Just before use, 35 µl of β-mercaptoethanol was added per 10 mL of lacZ-buffer.

Method:

The secondary growth cultures of strains harbouring pSD5B derivative promoter probe plasmids were grown in 7H9 broth containing OADC till the OD_{600} reached 0.8 - 0.9. 100 µl of each culture was added to 0.9 mL of Z-buffer (pre-heated at 37°C heat block) in a 2 ml Eppendorf tube. One drop each of 0.1% SDS and chloroform was added to the mixture. The tube was vortexed for 60-90 seconds to make the cells permeable to ONPG. To start the reaction 200 µl ONPG was added in the reaction mixture; recorded the time and allowed the reaction to develop yellow colour at 37°C. When sufficient yellow colour developed, the reaction was stopped by addition of 0.5 ml of 1M Na_2CO_3. The reaction time was noted down. The reaction mixture was centrifuged at high speed and OD_{420} and OD_{550} of the yellow supernatant was measured against Z buffer as a blank. Promoter activity was calculated in miller units according to the following formula.

$$\text{Miller Units} = 1000\ X\ \frac{O.D420\ -\ (1.75\ X\ O.D550)}{O.D600\ X Culture\ vol.(mL)\ X\ Time\ of\ incubation\ (minutes)}$$

2.5.3 Complementation assay

The 5 ml overnight growth cultures of *E. coli* strains harbouring pTrc derivative plasmids listed in Table 2.2 were pelleted down and supernatant was removed. The pellet was resuspended in 5 ml of fresh LB broth. Optical density of all the cultures was measured and normalized. 10-fold serial dilutions were made from neat to 10^{-7}. 5 ml of each dilution was spotted on LB agar plates containing suitable additives and antibiotics. The spots were dried, and plates were incubated at suitable temperature.

2.6 General molecular biology techniques

2.6.1 Sanger sequencing

DNA sequences of recombinant plasmids were determined by Sanger sequencing (Sanger et al., 1977) SpeedREAD service provided by Source Biosciences. The primers were synthesized either by Eurogentec https://www.eurogentec.com/en/ or Sigma Aldrich. For sequencing, 100 ng/ml plasmid DNA and 3.2 pmol/ml of primers in a volume of 5 ml per reaction each were sent to Source Biosciences.

2.6.2 Plasmid cloning

2.6.2.1 Cloning by restriction digestion

Following steps were followed for cloning of fragments purified by restriction digestion.

1. Restriction digestion

Both the plasmids (backbone and the source of fragments to be cloned in the backbone plasmid) were digested by appropriate restriction enzymes according to the parameters given in Table 2.15.

Component	Reaction volume for X µg DNA	Reaction volume for 1 µg DNA
Plasmid DNA	5 to 10 µg	1 µg
10X Buffer	1 X	5 µl
Restriction enzyme (s)	5 to 10 units (1 unit per µg DNA)	1 µl
Nuclease free water to make total volume	50 µl/ µg DNA	50 µl

Table 2.15 Restriction digestion parameters.

All the restriction digestion reactions were performed by using NEB or Thermo Fisher Scientific restriction enzymes with the recommended reaction buffers of the company. All the restriction digestion reactions were incubated for 1 hour at the optimum temperature for the enzyme.

2. Gel purification

After restriction digestion the reactions were run on 0.8 – 1.0% agarose gel. The gel purification of the required fragments was performed using the QIAquick Gel Extraction Kit (QIAGEN) following the manufacturer's recommended protocol. Briefly, the gels were excised along with the required fragments and transferred in an Eppendorf tube. The weight of the excised gel fragments was measured. Three volume Buffer QG was added to 1 volume of the gel and heated at 50°C heat block for 10 minutes (vortexed every 3 minutes to ensure the complete mixing of the gel). One gel volume isopropanol was added to the sample and mixed by inverting the tube. The sample was then applied to QIA quick column, spun down and discarded the flow

through. 500 ml buffer QG was added to the column and centrifuged for 20 seconds. The column was washed with 750 ml buffer PE and finally, the DNA fragment was eluted in buffer EB. All the centrifugation steps were done at 13K rpm.

3. Ligation

The fragments were ligated in the plasmid backbone using 10:1 molar ratio of insert to plasmid using T4 DNA ligase provided by NEB. All the ligation reactions were incubated overnight at 16-18°C.

4. Transformation and screening

All the ligation reactions were transformed in ultracompetent TOP10 cells. The colonies were screened using colony PCR or by purification of plasmids followed by restriction digestion. All the plasmid constructs were confirmed by sequencing.

2.6.2.2 Cloning of PCR fragment

The upstream fragments of *cpn* genes of 1218R were amplified by PCR from genomic DNA for cloning into pUC18 using blunt end ligation. *cpn10* gene was also amplified using the same method to clone into pTrc99A. 50 ml PCR reaction was set up according to the parameters given in Table 2.16. Following PCR, the PCR product clean-up was performed using QIAquick PCR Purification Kit (QIAGEN) following the manufacturer's recommended protocol. The parameters used for PCR reaction are given in Table 2.17.

ddH_2O	33 µl
Genomic DNA	2.0 µl
Forward Primer (10 µM)	2.0 µl
Reverse Primer (10 µM)	2.0 µl
10X pfu buffer with $MgSO_4$	5.0 µl
dNTP mix (2mM each)	5.0 µl
Pfu polymerase	1.0 µl

Table 2.16 PCR reaction set up for amplification of fragments from genomic DNA.

Segment	Step	Cycles	Temperature	Time
1	Initial denaturation	1	95°C	2 minutes
2	Denaturation	40	95°C	30 sec
	Annealing		50°C	30 sec
	Extension		72°C	1 minute (30 sec/kb)
3	Final extension	1	72°C	10 minutes

Table 2.17 Parameters used for PCR amplification.

2.6.2.3 Cloning of Oligos

Reagent required:

Tris buffered saline (TBS): 10mM Tris-HCl pH 8.0, 0.1mM EDTA, 150mM NaCl

Method:

For cloning of 31 bps oligos (Sph fragment), both the forward and reverse oligos tubes were spun down at 14k rpm for 2 minutes. 100 mM master stock was prepared by adding sterile water according to the data sheet provided by the manufacturer. The tubes were incubated at 42°C for 15 minutes and spun down at 7K rpm for 1 minutes. Both the oligos were then phosphorylated at 37°C for 1 hour using T4 PNK (NEB) enzyme according to the parameters given in Table 2.18.

Sph_Forward fragment		Sph_Reverse fragment	
Nuclease free water	15 µl	**Nuclease** free **water**	15 µl
Sph_F (From 100 µM Stock)	2 µl	Sph_R (From 100 µM Stock)	2 µl
10X T4 DNA Ligase Buffer	2 µl	10X T4 DNA Ligase Buffer	2 µl
10 mM ATP	2 µl	10 mM ATP	2 µl
T4 PNK	1 µl	T4 PNK	1 µl
Total	22 µl	Total	22 µl

Table 2.18 Reagents used for the phosphorylation of oligos.

Following phosphorylation, T4 PNK was heat inactivated at 65°C for 20 minutes. Both the oligos were annealed by adding 10 µl of each phosphorylated fragments in 80 µl of TBS in a microcentrifuge tube. The tube was kept on a heat block at 95°C for 3 minutes. The heating block was then switched off to let it cool gradually to 35°C. The tube was

then transferred to thermocycler and incubated at 25°C for 30 minutes. The annealed fragments were then digested with SphI restriction enzyme and cloned into pSD5B derivative plasmid.

2.6.3 Site Directed Mutagenesis (SDM)

QuickChange Lightning Site Directed Mutagenesis Kit from Agilent Technologies was used for insertion or removal of a restriction site. For the synthesis of mutant strand PCR reaction was set up according to Table 2.19. PCR parameters for SDM are given in Table 2.20.

10x reaction buffer	5 µl
Plasmid DNA	2 µl
F_primer	1.25 µl
R_primer	1.25 µl
dNTP mix	1.0 µl
Quick solution reagent	1.5 µl
ddH_2O	38 µl
QuikChange Lightning Enzyme	1.0 µl

Table 2.19 PCR reaction set up for SDM.

Segment	Cycles	Temperature	Time
1	1	95°C	2 minutes
2	18	95°C	20 sec.
		60°C	10 sec.
		68°C	30 seconds/kb (3.2 minutes in this reaction)
3	1	68°C	5 minutes

Table 2.20 PCR parameters for site-directed mutagenesis.

Following PCR, 2 µl of Dpn1 enzyme was added to each amplification reaction, mixed thoroughly by pipetting, and incubated at 37°C for 5 minutes to digest the parental copies of the plasmids. The Dpn1-treated products were then transformed into XL10-Gold Ultracompetent cells according to the manufacturer's protocol and plated out onto LBA plates containing the appropriate antibiotics.

2.7 Protein Analysis techniques

2.7.1 Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE)

Reagents required:

1. Double distilled water (ddH_2O)
2. 30% acrylamide/Bis mix (Bio-Rad)
3. 1.5 M Tris pH 8.8 (for resolving gel): 18.15 g Tris-base was dissolved in 80 ml of ddH_2O and pH was adjusted to 8.8 using HCl. ddH_2O was added to make the final volume of 100 ml.
4. 1.0 M Tris pH 6.8 (for stacking gel): 6 g of Tris-base was dissolved in 80 ml ddH_2O, pH was adjusted to 6.8 using HCl and the final volume was made up to 100 ml with ddH_2O.
5. 10% SDS
6. 10% Ammonium persulfate (APS)
7. TEMED (Sigma-Aldrich)
8. SDS Loading buffer 5X: Loading buffer was made by mixing the reagents given in Table 2.21 and adjusting the final volume to 10 ml with ddH_2O. β-mercaptoethanol was added just before use.

Reagent	Quantity used for 10 ml	Final concentration
1.0 M Tris-HCl, pH 6.8	2.5 ml	250 mM
SDS	1 g	10%
Bromophenol blue	1 mg	0.1%
Glycerol 100%	4 ml	40%
β-mercaptoethanol (100% stock)	0.5 ml	5%

Table 2.21 Reagents used for the preparation of SDS 5X Loading buffer.

9. Running buffer 10X: 30.3 g Tris-base, 144.4 g glycine and 10 g SDS was dissolved thoroughly in 800 ml ddH_2O, and the final volume was adjusted to 1 L with ddH_2O.
10. Staining solution: 0.25 g of Coomassie Brilliant Blue R was added in 50 ml of methanol and 10 ml of glacial acetic acid. The total volume was adjusted to 100 ml with ddH_2O.
11. Destaining solution: Destaining solution was prepared by mixing of 50 ml methanol (50%), 10 ml glacial acetic acid (10%) and 40 ml of ddH_2O (40%).

12. 10 M urea (for urea lysis of the sample)

Method:

1. Sample preparation:

5 µl of each of the secondary growth cultures was inoculated with 10 ml of primary growth cultures and incubated in a shaking incubator at 30°C until the OD_{600} reached to ~ 0.5. The cultures were then induced with 0.1 mM IPTG and continued to grow until the OD_{600} reached to 1.0. 1 ml of each of the secondary growth culture was transferred into 1.5 ml Eppendorf tube and centrifuged at 7000 rpm for 10 minutes. The supernatant was removed carefully, and the pellet was resuspended in 70 ml of ddH_2O, 30 µl of 10 M urea and 20 µl of 5X SDS Loading buffer. The samples were boiled for 15 minutes at a heating block, spun at high speed. Negative and positive controls were always included in the samples.

2. Gel preparation:

Resolving and stacking gels were prepared according to the recipe given in Table 2.22.

10% Resolving gel			Reagents	5% Stacking gel		
30 ml	20 ml	15 ml		4 ml	5 ml	8 ml
11.9	7.9	5.9	ddH_2O	2.7	3.4	5.5
10.0	6.7	5.0	30% acrylamide/Bis mix	0.67	0.83	1.3
7.5	5.0	3.8	1.5 M Tris pH 8.8	-	-	-
-	-	-	1.0 M Tris pH 6.8	0.5	0.63	1.0
0.3	0.2	0.15	10% SDS	0.04	0.05	0.08
0.3	0.2	0.15	10% Ammonium persulfate	0.04	0.05	0.08
0.012	0.008	0.006	TEMED	0.004	0.005	0.008

Table 2.22 Reagents used for the preparation of SDS-PAGE gel.

3. Gel Running

The gel was assembled in the Bio-Rad mini-gel system according to the manufacturer's instructions. The tank was filled with 1X running buffer and wells were cleaned by flushing with running buffer to clean any unpolymerized acrylamide. 20 ml of each of the sample was loaded into each well, in one well pre-stained broad range protein

ladder (NEB) was also loaded, and the gel was subjected to electrophoresis at 100 V for 90-120 minutes.

4. Staining and de-staining

The gel was removed from the tank after electrophoresis and stained using Coomassie Brilliant staining solution for about 2-3 hours on a low-speed blot shaker. The stained gel was de-stained by repeatedly by washing the gel with de-staining solution.

2.7.2 Native-PAGE

Reagents required:

1. Lying buffer: 50 mM Tris (pH 7.4), 150 mM NaCl, 1 mM EDTA. Two protease inhibitor tablets were added to the buffer before use.
2. 1 X running buffer: 3.0 g Tris-base and 14.4 g glycine was dissolved thoroughly in 800 ml ddH_2O, and the final volume was adjusted to 1 L with ddH_2O.
3. 4X native gel loading dye: 0.25 M TRIS/HCl (pH 6.8), 80% glycerol, 0.04% bromophenol blue.

Method:

50 ml secondary growth culture was pelleted down, supernatant was removed, and pellet was resuspended in 0.5 ml lysing buffer. The cell solution was transferred to Lysing Matrix E tube and cells were lysed using FastPrep MP Biomedicals. The tubes were spun down at 13K rpm for 15 minutes at 4°C. Native page sample was prepared by mixing 50 ml 4X native gel loading dye with 150 ml of cell on ice. The gel was run in a tank filled with 1 X native page running buffer at 150 V for 2 hours.

2.7.3 Western blotting

Materials and reagents required:

1. Pre-assembled PVDF membrane (Invitrogen)
2. TBST 10X (stock): 10X TBST was prepared by adding 24.23 g Tris-HCl (or Tris-base) and 88 g of NaCl to 800 ml distilled water. Adjusted the pH to 7.6 by adding HCl or NaOH as required. The final volume was adjusted to 1 L by adding ddH_2O. 10 ml (1%) tween 20 was also added to the solution.
3. Blocking solution 5%: 1 g fat-free milk was added to 20 ml of 1X TBST.
4. Primary antibody: Monoclonal IgG anti-rabbit primary antibody specifically binding to the C-terminal of MmCpn60.1.
5. Secondary antibody: Anti-rabbit IgG secondary antibody.
6. BCIP/NBT- purple liquid substrate (Sigma).

Method:

1. Transfer of proteins to PVDF membrane

Invitrogen pre-assembled transfer stacks with integrated PVDF membrane was used to transfer the proteins from SDS-PAGE or Native page to PVDF membrane. The transfer was done using Invitrogen iBlot 2 dry blotting device. iBlot 2 stacks consists of a copper cathode and anode which sandwich dry gel matrices containing appropriate buffers. The transfer stack was unsealed, top stack was separated with the help of tweezers whereas bottom stock was left in the plastic tray (the bottom stack contains the PVDF membrane). Invitrogen iBlot 2 device was switched on and the tray was placed on the blotting surface by aligning the electric contexts of the try with the correspond contexts of the iBlot surface. The pre-run SDS-PAGE or Native-PAGE was placed on the PVDF membrane. Next, an iBlot filter paper (pre-soaked in ddH_2O) was placed on top of the gel. The bubbles were removed using the roller. The top stack was placed over the filter paper and the bubbles were again removed using the roller. The absorbent pad was placed on top of the top stack by aligning the electrical contexts. The lid of the iBlot was closed and pre-programmed template (P3; 20V, 0.9A, 7min) was used to run the transfer. In this way the proteins were completely transferred from the gel to the blotting membrane.

2. Blocking

The top stack was carefully removed from the tray and PVDF membrane was transferred to 20 ml of the 5% blocking solution in a clean square petri plate. The plate was kept overnight on low-speed blot shaker.

3. Blotting

Following morning 10 ml of primary antibody was added to the blocking solution and incubated for 1 hour on the shaker at room temperature. The PVDF membrane was washed 3 times (5 minutes each) with 15 ml of 1X TBST. 1 µl of the secondary antibody was added to fresh 20 ml of 5% blocking solution. The membrane was transferred to the blocking solution, incubated for 30 minutes, and washed 3 times (5 minutes each) with 15 ml 1X TBST.

4. Detection

NBT/BCIP purple liquid substrate is a one-component sensitive precipitating alkaline phosphatase substrate used for the detection of the bands.

2.7.4 Protein purification

Two methods were used for the purification of *M. marinum* Cpn60.1.

2.7.4.1 Nickle-nitrilotriacetic acid Ni-NTA column protein purification

Material and reagents required:

1. Tris (pH: 8.0) 1M solution (500 mL)
2. NaCl 5M solution (100 mL)
3. Imidazole 1M solution (500 mL)
4. Lysis/Equilibrium buffer: 100 ml of lysis/equilibrium buffer was prepared by adding 50mM Tris and 150 mM NaCl. Two Protease inhibitor tablets were added before use.
5. Wash buffers: 30 ml of wash buffers containing 50mM Tris and 150 mM NaCl with varying concentration (5 mM, 10 mM, 15 mM, 20 mM, 25 mM, 50mM) of imidazole were prepared. Wash buffers were named as W_1, W_2, W_3, W_4, W_5, W_6 respectively.

6. Elution buffers: 30 ml of elution buffers containing 50mM Tris and 150 mM NaCl with varying concentration (100 mM, 150 mM, 200 mM, 250 mM) of imidazole were prepared. Elution buffers were named as E_1, E_2, E_3, and E_4 respectively.
7. Ni-NTA column

Method:

1. Sample lysis:

1 L of the secondary growth culture of *M. marinum* 1218R W.T strain was centrifuged at 7000 rpm for 10 minutes. The supernatant was removed, and the pellet was suspended in 1.5 ml of lysis buffer. The mixture was transferred to the lysing matrix E tube (MP Biomedicals, 2 ml tube) and the cells were homogenized using the FastPrep MP biomedical instrument. The lysed sample was centrifuged at 12K rpm for 20 minutes. The supernatant (Load, L) was ready to load onto the column.

2. Column equilibration and sample loading:

600 ml of equilibration buffer was added to Ni-NTA column, the column was spun at low speed (2K rpm) for 20 seconds. The process was repeated 10 times to equilibrate the column. The supernatant (L) was applied onto the column and spun down at 2K rpm for 20 seconds. The His-tag protein was now bound to the resin. The flowthrough (F) was saved in an Eppendorf tube.

3. Washing and elution:

The column was washed one by one with wash buffers. 600 ml wash buffer was added to the column and the column was centrifuged at 2K rpm for 20 seconds. Flow through was collected and named as W_1, W_2, W_3, W_4, W_5, W_6. To find out the optimized concentration of imidazole in the elution buffer, all the four elution buffers were used one by one in the first attempt. 200 ml of each elution buffer was added to the column and flow through was collected by spinning the column at 2K rpm for 20 seconds. The elations were named as E_1, E_2, E_3, and E_4. 20 ml of load (L), initial flowthrough (F), wash flow through (W_1, W_2, W_3, W_4, W_5, W_6) and elution flow through (E_1, E_2, E_3, and E_4) were loaded on SDS-PAGE gel. Western blotting was also performed to confirm the desired protein.

2.7.4.2 Immunoprecipitation

Material and reagents required:

1. Lysing buffer/ Binding buffer: Phosphate buffered saline (PBS), pH 7.4. Protease inhibitor tablets were added just before use.

2. Elution buffer 100 mM glycine-HCl, pH 2.5-3.0: Prepared 0.1 M glycine solution and adjusted the pH to 2.8 using HCl.

3. Primary antibody: Anti-rabbit IgG antibody specific to the protein of interest
4. Protein G Mag Sepharose™ Xtra beads (Cytiva)

Method:

1. **Bacterial culture**
   One litre secondary growth culture of *M. marinum* 1218R grown at 30°C in mid log phase (pH 0.8-1.0) was centrifuged to harvest the pellet. The pellet was kept on ice. All the subsequent steps were done on ice unless stated otherwise.
2. **Preparation of Cell lysate**

Two protease inhibitor tablets were added to the lysis buffer (PBS). 2 ml of the lysis buffer was added to the cell lysate, mixed well by pipetting, and added to the Lysing matrix E tubes. The cells were lysed using FastPrep instrument MP Biomedicals according to the manufacturer's instructions. The lysed sample was centrifuged at high speed (12 K rpm) for 20 minutes. Supernatant (cell lysate) was kept on ice in a sterile Eppendorf tube.

3. **Equilibration of magnetic beads (for pre-cleaning of cell lysate)**

The magnetic beads medium slurry was mixed thoroughly by vortexing. 50 ml of the homogenous bead slurry was taken into an Eppendorf tube. Placed the Eppendorf tube in the magnetic rack, the pellet of the beads was formed separating the storage solution. The storage solution was removed completely by pipetting. The beads were washed twice with 500 ml of lysis buffer. For washing the beads lysis buffer was added to the bead pellet, vortexed briefly, placed back in magnetic separation rack and the buffer was removed by pipetting.

4. Cell lysate pre-clearing

1ml of cell lysate was added to 50 µl of magnetic beads (pre-washed) and incubated for 20 minutes at room temperature on a benchtop rotator. The beads were separated from the lysate by pelleting down using the magnetic rack. The pre-cleaned cell lysate was transferred to a new sterile Eppendorf tube. The bead pellet was discarded.

5. Preparation of immunocomplex

30 µl of primary antibody was added to the pre-cleaned cell lysate and incubated for 8 hours at 4°C on a benchtop rotator to prepare the immunocomplex.

6. Equilibration of magnetic beads

100 µl of the homogenous magnetic bead slurry was taken into an Eppendorf tube. Beads were pelleted down by placing the tube in the magnetic rack. Storage solution was removed, and beads were washed twice with 1 ml of lysis buffer.

7. Sample application

The immunocomplex solution was immediately transferred to the pre-washed magnetic bead pellet tube. The tube was incubated at room temperature for 30 minutes on a slow rotatory shaker.

8. Washing

After 30 minutes, beads were pelleted down using magnetic rack. The beads were washed 5 times with 500 µl binding buffer. The tube was kept on ice between washes.

9. Elution

600 µl of elution buffer was added to the pellet after washing. The beads were thoroughly resuspended in the buffer followed by the separation of beads using magnetic rack. The elution fraction (containing the immunocomplex) was collected by pipet and transferred to a sterile Eppendorf tube.

10. Cold acetone precipitation

The eluate was transferred to a 50 ml falcon tube and acetone was added four times the volume of the sample. The tube was kept at -20°C overnight for precipitation of the protein. The flowing morning, the tube was spun down at high speed, acetone was removed, and the pellet of the protein of interest was dissolved in 100 µl of 0.5M urea solution. 10 µl of the sample was used for analysis by western blotting and the rest of the sample was used for mass spectrometry analysis.

2.8 Bioinformatics Analysis

2.8.1 Genome browsers

All the genome sequences were downloaded from NCBI using accession numbers given in table 3.1. Visual exploration and comparison of genomes was done using Artemis https://www.sanger.ac.uk/tool/artemis/, Artemis comparison tool ACT (Carver et al., 2005), IGV (J. T. Robinson et al., 2011) or Mauve genome alignment tool (Darling et al., 2004).

2.8.2 Phylogenetic tree construction

Maximum likelihood phylogenetic trees were constructed with 1000 bootstraps using default parameters in MEGA X or MEGA 11 software.

2.8.3 Roary Analysis

The stand-alone pan-genome pipeline, Roary, was used to determine the genes presence/absence in the genomes. The pipeline was installed from https://sanger-pathogens.github.io/Roary/. PROKKA annotation files or NCBI GFF3 files were used for Roary analysis according to the requirements of the analysis. For running the pipeline, default parameters were used.

2.8.4 RNA-seq Analysis

For RNA seq analysis, FASTQC was used for quality control of the FASTQ reads files. The reads were aligned to the reference genome using HISAT2 software. The output SAM files were converted to BAM files and indexed using SAMtools. Reads were quantified using HTSeq-count and finally the differentially expressed genes were detected using DESeq2.

3 Genome comparison and phylogenetic relationship

3.1 Introduction

It was stated in the introduction chapter (section 1.2.2) that *M. marinum* is a newly developed model organism being used increasingly to study the pathogenesis of *M. tuberculosis*. One way to get an insight if an organism is a good model to be used for the study of the pathogenesis of another organism is to compare their genomic sequences and to find out the phylogenetic relationships between them. Comparative genome study helps to identify the core (conserved) genes and tells how closely the two species are related to each other. It is also a powerful tool to study the evolutionary relationships among species. A Phylogenetic tree provides understanding of the evolution of different species and their relatedness to each other (Munjal et al., 2019). What follows is a brief review of the published studies on comparative genome and phylogenetic relationship of *M. marinum* and *M. tuberculosis*.

M. tuberculosis and *M. marinum* are close genetic relatives (Stinear et al., 2008). Several studies have shown a close phylogenetic relationship of >85% nucleotide identity between the two species (Devulder et al., 2005; Gey van Pittius et al., 2006a; Yip et al., 2007). The nucleotide identity of *M. marinum* was found to be >95% with *M. ulcerans* and >85% with *M. tuberculosis*. Studies of the whole genome comparison revealed that 80% of all the coding DNA sequence (CDS) of *M. tuberculosis* (H37Rv) show sequence homology with *M. marinum* Musa (M strain) with an average of 85% amino acid identity (Stinear et al., 2008).

Phylogenetic analysis based on the alignment of partial gene sequences of 16S rRNA of 22 mycobacterial species showed that *M. mairnum* and *M. ulcerans* are more closely related than any other species. Both organisms were also observed to be more closely related to the *M. tuberculosis* than many other species outside the *M. tuberculosis* complex (Prasanna & Mehra, 2013; Tønjum et al., 1998). In another study, the maximum likelihood tree based on the 1941 core protein sequences of 150 mycobacterial species showed a similar close phylogenetic relation between *M. marinum* and *M. tuberculosis* (Gupta et al., 2018).

In another study complete genomes of 157 mycobacterial species were compared to find out the core genes. 304 single copy genes present among all the 157 species were then used to construct the phylogenetic tree. The tree showed that the ancestral strains

are more closely related to today's fast-growing mycobacteria while the slow-growing species make a separate arm. *M. smegmatis* (a fast-growing mycobacterium) is less closely related to *M. tuberculosis* while *M. marinum* shares a more recent common ancestor with *M. tuberculosis* (Bachmann et al., 2020) (Figure 3.1).

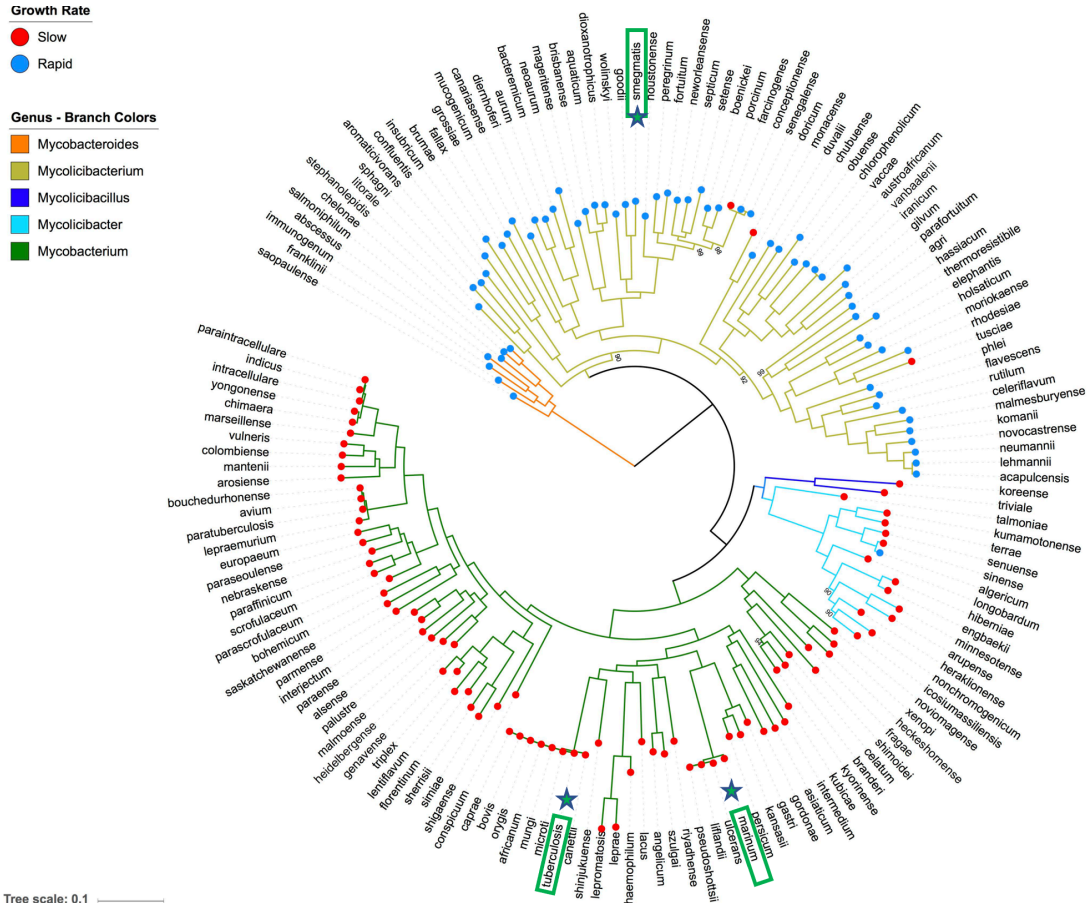


Figure 3.1 Maximum likelihood tree based on 304 core genes of 157 mycobacterial species. Image taken from (Bachmann et al., 2020). *M. marinum*, *M. tuberculosis* and *M. smegmatis* are boxed.

Considerable genomic diversity was also observed among different strains of *M. marinum*. In an extensive study with 17 different strains of *M. marinum* based on genome comparison, average nucleotide identity and phylogenetic analysis showed the division of *M. marinum* strains in two distinct clusters; cluster I comprised of six strains including Musa while cluster II included mainly E11, 1218R, and their derivative strain (Das et al., 2018). In another study the genome comparison of the three strains of *M. marinum* namely M, E11, and ATCC 927 showed that all the three strains share 4330 orthologous gene clusters but also possess strain specific gene clusters: 287, 714 and 1223 in E11, M, and ATCC 927 respectively (Yoshida et al., 2018).

No previous studies on the whole genome comparison of *M. marinum* 1218R and *M. tuberculosis* are available in the published literature.

Besides, both the homologous chaperonins (i.e., Cpn60.1 and Cpn60.2) of mycobacteria have high sequence conservation and amino acid identity yet have unique regions of dissimilarity and are potentially involved in various moonlighting activities (Henderson et al., 2010; Sielaff et al., 2011). Phylogenetic analysis based on the nucleotide or protein sequences of the chaperonins could suggest useful hypotheses about the emergence of multiple chaperonin genes in actinobacteria as well as the evolutionary relationship and relatedness of the species.

In this study, I used *M. marinum* 1218R strain to investigate the mechanism and function of chaperonins of *M. marinum*. To relate the findings of this study with the *M. tuberculosis* H37Rv, it is important to investigate the genome comparison of the two strains. Therefore, in the following sections of this chapter, the genome comparison of the two species will be discussed briefly for the better understanding of the relationship of the two species and the phylogenetic analysis of various species of mycobacterial strains along with some other actinobacteria will be analysed. Finally, to find out the conserved and variable regions of both the chaperonins, the amino acid sequence alignment of the chaperonins will also be examined.

3.2 Analysis plan

For the comparison of the whole genomes, FASTA, GFF3 and Genbank files of *M. marinum* 1218R and *M. tuberculosis* H37Rv were downloaded from NCBI (https://www.ncbi.nlm.nih.gov). The genomes were compared and analysed using Artemis (http://sanger-pathogens.github.io/Artemis/Artemis/), Roary (https://sanger-pathogens.github.io/Roary/), and Mauve (https://darlinglab.org/mauve/mauve.html) software.

To investigate the species relatedness, evolution, and phylogenetic relationship 50 species of the genus mycobacteria and 12 species of other actinobacteria were included in the analysis. *E. coli*, *Staphylococcus aureus,* and *Bacillus subtilis* were used as outgroup organisms in this analysis. Phylogenetic analysis was performed separately based on the nucleotide sequences of 16S rRNA gene and the hsp60 chaperonin homologous genes using MEGAX or MEGA11 software (https://www.megasoftware.net). Multiple sequence alignment of all the chaperonin homologs was also performed, to investigate the sequence identity and the variable and conserved regions, either in MEGAX or DNASTAR MegAlign Pro (https://www.dnastar.com/software/lasergene/megalign-pro/).

It was hypothesized that *M. marinum* 1218R would show considerably close genomic and phylogenetic relationship with *M. tuberculosis* H37Rv providing the basis to be considered as a promising model to study the *M. tuberculosis* pathogenesis.

3.3 Results

3.3.1 Whole genome comparison of *M. marinum* and *M. tuberculosis*

The complete genome sequences (FASTA and GenBank formats) of the *M. marinum* 1218R (strain used in this study) and *M. tuberculosis* H37Rv were downloaded from NCBI. *M. marinum* (Mm) 1218R genome is comprised of a single circular chromosome of 6,467,358 bps with an overall GC content of 65.7% and a circular plasmid of 127,402 bps called pMMRN with an overall GC content of 63.93%. *M. tuberculosis* on the other hand has a smaller genome consisting of a linear chromosome of 4,411,532 bps with an overall GC content of 65.61%.

The sequence and annotation features of both the genomes were compared and analysed to find out the similarities and relatedness of the two organisms. The chromosome of *M. marinum* 1218R comprises 5405 coding DNA sequence (CDS) features including 5213 genes (covering 86.7% of the coding region) and 192 pseudogenes and the plasmid pMMRN sequence consists of 130 CDS. The genome of *M. tuberculosis* H37Rv contains 4032 CDS which includes 4018 genes (covering 91.2% of the coding region) and 13 pseudo genes.

3.3.1.1 Hypothetical proteins and PE/PPE family proteins

Searching through the whole genomes, using the Artemis genome browser, it was found that 1218R genome annotation contains 691 genes encoding hypothetical proteins which make 6.8% of the coding region and H37Rv genome has 713 of those which are 11.3% of the coding region. There is another group of proteins of unknown function called PE and PPE family proteins which are found abundantly in mycobacteria. These proteins are named so because of their characteristic motif of proline-glutamate (PE) or proline-proline-glutamate (PPE) at the positions 8, 9 and 8-10 of the N-terminus. H37Rv genome was found to have 166 genes encoding PE and PPE family proteins whereas 1218R contains 237 such genes which makes 6.4% and 7.3% of the coding region respectively. Studies have shown that these poorly understood proteins have significant importance in the mycobacteria; they might have coevolved with ESX loci, specifically ESAT-6 gene cluster, and expanded specifically. Multiple sequence alignment of these proteins showed extensive divergence at the C-

terminal region based on which these proteins can be subdivided in subfamilies suggesting the constant disruptive selection events occurred during evolution of these proteins (Garnier et al., 2003; Gey van Pittius et al., 2006b). This example shows that a considerable proportion of both the genomes is not well studied yet.

3.3.1.2 Other specific enzymes

Oxidoreductases are very important group of enzymes which play essential roles in many metabolic pathways including glycolysis, aerobic and anaerobic respiration. In *M. tuberculosis* they are also reported to be involved in cell homeostasis and antimicrobial resistance (Bhargavi et al., 2022). The multicopper oxidase might prevents the bacillus from ROS attack during infection besides its primary role of copper resistance. Triacylglycerol hydrolase (a PE/PPE domain containing protein of the known activity) was found to be involved in immunogenic activity (Kinkar et al., 2019; Mishra et al., 2008).

Searching through the genome of 1218R, 353 reductases and oxidoreductases genes, 189 genes for dehydrogenases, 34 oxidases and peroxidases genes, 218 CDS for hydrolases and 49 oxygenase enzymes genes were found which encompasses 18.2% of the coding region. H37Rv genome was found to have 180 reductases and oxidoreductase, 118 dehydrogenases, 29 oxidases and peroxidases, 77 hydrolases and 21 oxygenase enzymes genes making 12.35% of its coding capacity.

Glycotransferases are essential enzymes for the biosynthesis of mycolic acid cell wall of *M. tuberculosis* which in turn is necessary for the survival of the bacilli within the host (Berg et al., 2007). 1218R genome was found to code for 443 transferase enzymes including glycotransferases and acyltransferases comprising 7.5% of the entire coding capacity whereas H37Rv codes for 273 transferases making 7.0% of its coding capacity (CC).

Mm 1218R genome encodes 89 kinases (1.8% CC), 89 ligases (2.1% CC), and 19 lipases and phospholipases (0.3% CC). H37Rv was found to encode 70 kinases (1.9% CC), 75 ligases (2.7% CC) and 17 lipase/esterase enzymes (0.4% CC). 1218R genome contains 44 ATPase enzymes (1.2% CC) and H37Rv has 22 ATPases (1.0% CC).

3.3.1.3 Transport proteins and other secretion system proteins

Transport and secretion system proteins play crucial roles in the pathogenesis of *M. tuberculosis* (*Mtb*). For example, Mycobacterial membrane protein large (MmpL) proteins are involved in the transport of essential substrates of the cell wall of *Mtb* which in turn are substantial in the pathogenesis of Mtb (Melly & Purdy, 2019). In addition, ATP-binding cassette (ABC) transporter proteins are essential membrane proteins of mycobacteria which are involved in import and export of various fundamental substrates. Many ABC transporters were found to be involved in fatty acid transport and were suggested to be a possible target for tuberculosis treatment (Cassio Barreto de Oliveira & Balan, 2020).

1218R chromosome was found to contain coding sequences for 256 transporter proteins (4.9% CC) including 105 ABC transporters (1.8% CC), 32 Mmp family transporters (0.7% CC) and 43 MFS transporters (1.0% CC) whereas, H37Rv contain 146 transporter proteins genes (1.9%) which include 75 ABC transport proteins (1.9% CC) and 19 Mmp transporters (0.8% CC). H37Rv genome was not found to have any genes for MFS transporter protein.

Type VII secretion systems are fundamental for the persistence and pathogenesis of *Mtb* (Roy et al., 2020); Mm 1218R has 73 genes for such proteins (1.0% CC) and Mtb H37Rv contains 38 (1.3% CC) genes of secretion system proteins which as mainly ESX-1 to ESX-5 secretion system proteins.

3.3.1.4 Lipoproteins, transcriptional regulatory proteins, toxins/anti-toxins, and virulence proteins

Several lipoproteins and transcriptional regulatory proteins are known to be involved in the survival and virulence of *Mtb* during infection and various stress conditions (Becker & Sander, 2016; Raman et al., 2004). 1218R has 23 CDS for lipoproteins (0.2% CC) and 250 genes for transcriptional regulatory proteins (2.7% CC) whereas H37Rv possess 89 CDS for lipoproteins (1.7% CC) and 163 transcriptional regulatory protein genes (2.9% CC). In addition, H37Rv has surprisingly high number of toxin/antitoxin protein genes i.e., 123 (0.8% CC); 1218R contains only 10 genes coding for toxins/anti-toxins but 15 for virulence proteins (0.2% CC).

3.3.1.5 Plasmid pMMRN

M. marinum (Mm) 1218R plasmid sequence pMMRN was analysed. It consists of 130 coding DNA sequences (CDS) which make 73.8% of the plasmid sequence. Out of 130 CDS, 110 annotations are hypothetical proteins (59.5% CC) whereas the remaining 20 CDS encode for secretion system associated proteins (Type VII, Esx-5), ATPase and ATP binding proteins, transcriptional regulatory protein and some are enzymes like integrase, prolin imino-peptidase, methanol dehydrogenase, transposase, and a toxin Doc protein.

3.3.1.6 Construction of genome maps of 1218R and H37Rv

All the CDS features described in the section 3.3.1.1 – 3.3.1.5 were colour coded in the Artemis genome browser and the maps of the genomes and plasmid were created in Artemis DNA plotter (Figure 3.2 (A & B) and Figure 3.3). GC skew (G − C)/(G + C) plots showed a bias in GC content between the two DNA strands.

3.3.1.7 Comparison of genomes using the Artemis comparison tool (ACT) and MAUVE alignment

To visualize the comparison of both the genomes (Mm 1218R and Mtb H37Rv), a genome comparison file was created. FASTA sequence files of Mm 1218R (subject) and Mtb H37Rv (query) were aligned in NCBI blastn using discontiguous megablast (recommended for cross-species comparison) and the resulted hit table (text) files were downloaded. The GenBank format genome files of Mm 1218R and Mtb H37Rv were visualised in ACT (Carver et al., 2005) along with the hit table (text) file created by alignment. The colinear and synteny regions of the two genomes are shown by red and blue lines in the display window (Figure 3.4 A). Red lines indicate the collinear regions on of both the genomes whereas blue lines show the collinear regions which had undergone inversion event in one of the two organisms. The intensity of red or blue line is directly proportional to the percentage identity of the two matches. For better understanding, the genome of *M. smegmatis* MC^2155 (another model organism used to study Mtb pathogenesis) was also compared with the genome of H37Rv (Figure 3.4 B). Segments of inversion (blue lines) were observed in both comparisons.

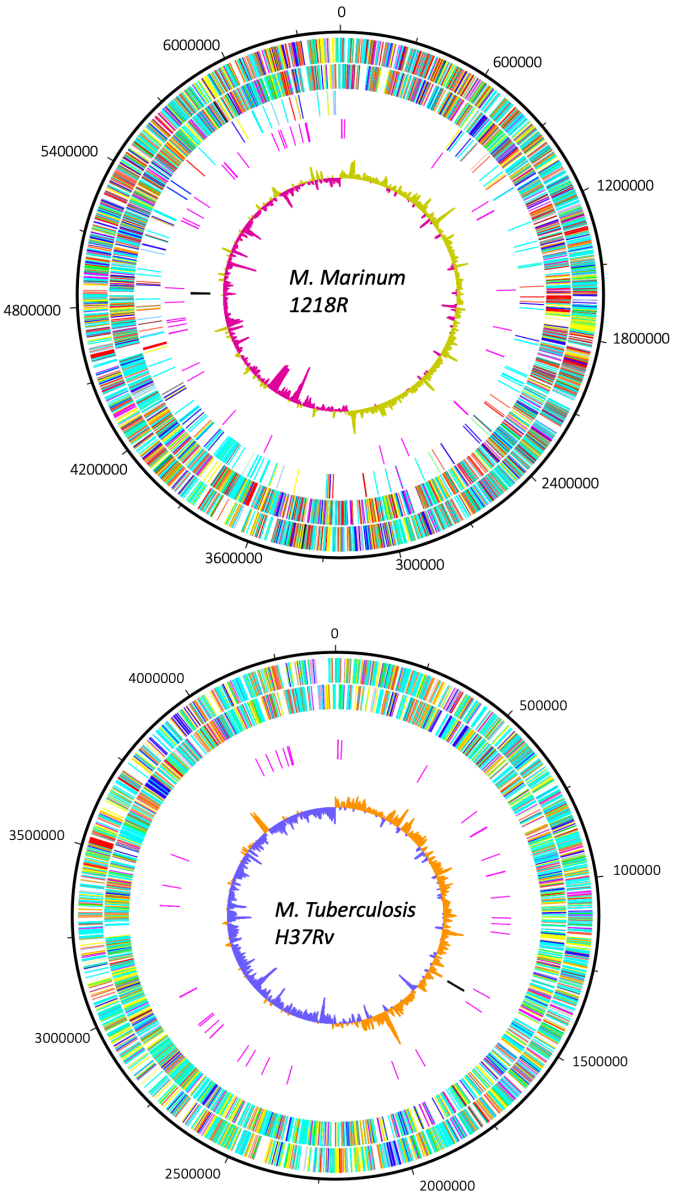


Figure 3.2 Graphical representation of the genome map of *M. marinum* 1218R and *M. tuberculosis* H37Rv.

A) Graphical representation of the genome map of *M. marinum* 1218R chromosome constructed using Artemis DNA plotter software. B) Circular map of M. tuberculosis H37Rv chromosome constructed using Artemis DNA plotter. The first (outermost) track represents all the CDS present on the forward strand and the second track represents all the CDS present on the reverse strand. All the pseudogenes are marked in the third track. Track four represents all the tRNA genes whereas rRNA gene cluster is shown in black colour in the next track. GC skew plot is shown in the middle of the map. Gene colour codes: PE, PPE, PE-PPE family genes (blue), hypothetical protein genes (yellow), genes for reductases (red), dehydrogenases (orange), transferases (tomato), oxidases and peroxidases (tomato), hydrolase (tomato), transcription associated genes (purple), ATPases (pink), kinases (green), ligases (brown), secretion system associated genes (plum), and ABC transporter family genes (Magenta).

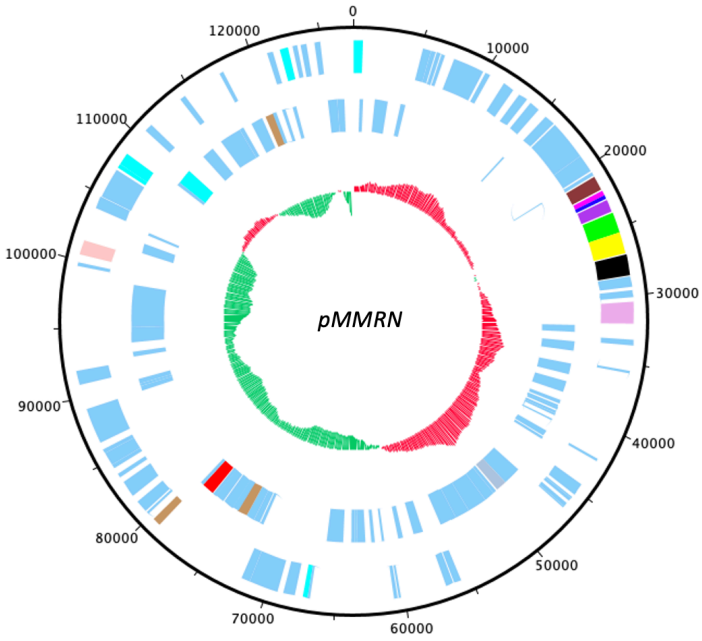


Figure 3.3 Graphical illustration of *M. marinum* 1218R plasmid pMMRN.
Map was constructed using Artemis DNA plotter. The first circle (outermost) encompasses all the CDS in the forward strand whereas the second displays all the CDS in the reverse strand. GC skew is plotted in the middle. Hypothetical protein gene (sky blue).

M. tuberculosis H37Rv and *M. marinum* 1218R showed homologous collinear regions at both ends of the chromosome and in the middle region whereas inversion regions were present between the collinear regions (Figure 3.4A). Similarly, MC²155 and H37Rv comparison showed collinear regions in the middle and in the start and end of the chromosome and inversion regions were present in between them (Figure 3.4 B).

To visualize the regions of inversion more clearly, the genomes of the three species (*M. marinum* 1218R, *M. tuberculosis* H37Rv, and *M. smegmatis* MC²155) were aligned using MauveAligner (https://darlinglab.org/mauve/download.html). MAUVE analysis provides the synteny of the genes in the genomes. As a result of MAUVE genome alignment, the conserved regions of the chromosome in all the species under examination are shown as localized collinear blocks (LCBs). Each LCB shows a segment of chromosome without any rearrangement of the original sequence obtained from the common ancestor. LCBs below the centre line of each genome represent the sequences present in the reverse orientation in relation to the reference genome (Darling et al., 2004). The LCBs generated by MAUVE alignment are shown in Figure 3.5. As seen in the ACT comparison, *M. tuberculosis* H37Rv exhibits two clear regions of inversion compared to the 1218R and MC²155. As the regions outside the LCBs show areas of

dissimilarities, it can be seen in the Figure 3.6 that 1218R genome possess less regions of dissimilarities with H37Rv compared to MC²155. In the figure 3.5 inside each LCB, the similarity profile of the genome sequence is plotted. The white areas represent the unaligned sequences probably because they are strain specific regions. 1218R possess less unaligned sequences than MC²155. Overall, the result show that 1218R genome contain more sequence homology with H37Rv than MC²155 genome.

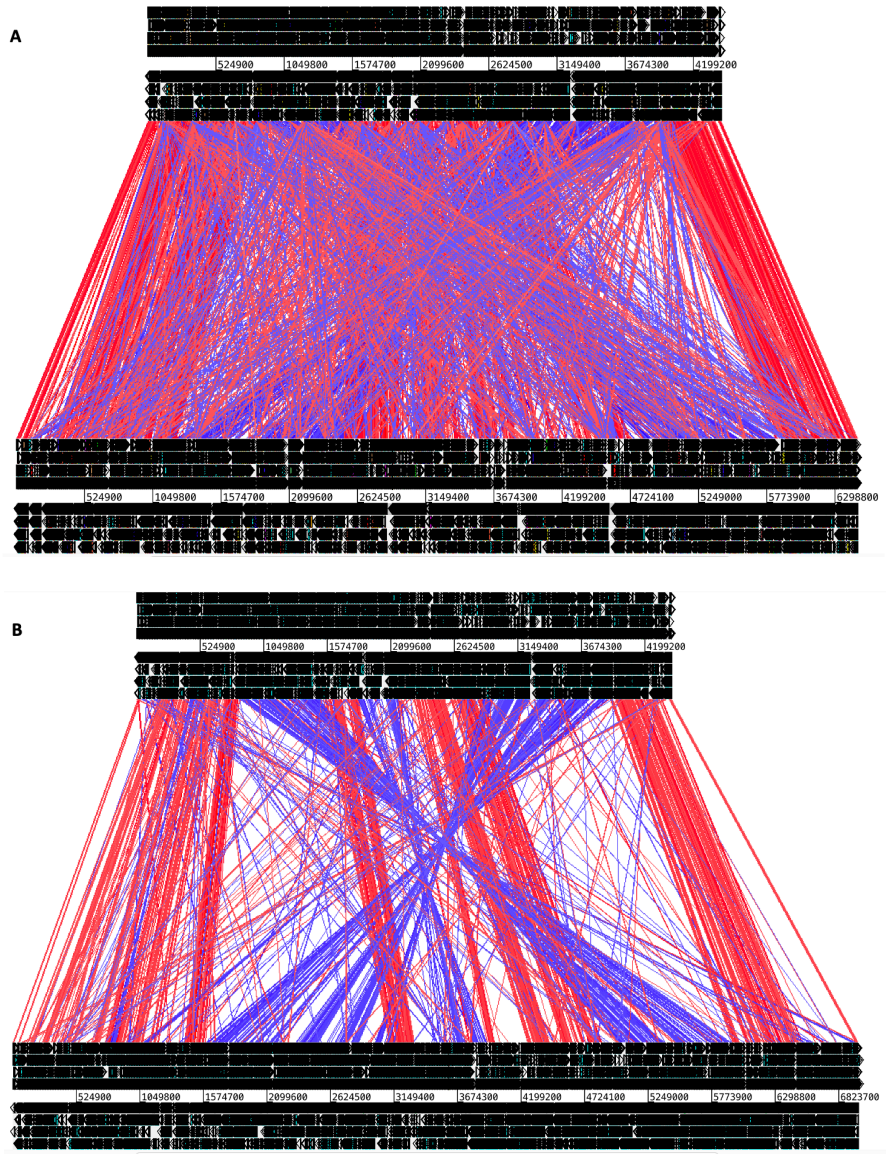


Figure 3.4 Genome comparison of *M. marinum* 1218R, *M. tuberculosis* H37Rv, and *M. smegmatis* MC2155 using ACT.
The display of the blastn comparison of two genomes A) H37Rv (top panel) vs 1218R (lower panel) and B) H37Rv (top panel) vs MC²155 (lower panel). The comparison view panel (middle panel) was dragged down to show only the most relevant regions.

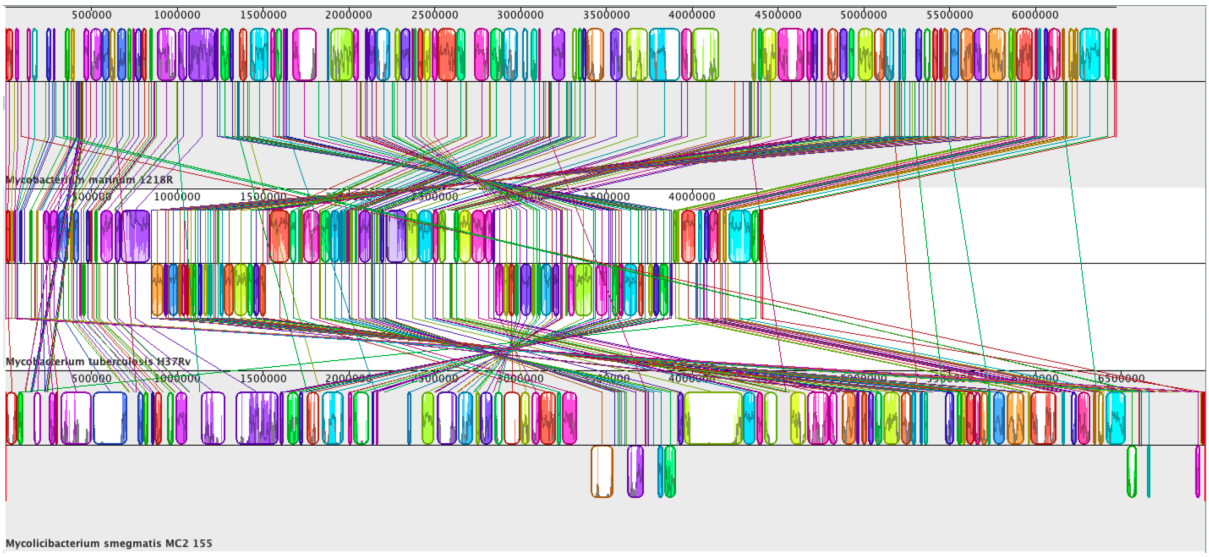


Figure 3.5 Genome alignment of mycobacterial genomes using MAUVE multiple alignment tool. From top to bottom strains are *M. marinum* 1218R (Top, reference), *M. uberculosis* tH37Rv (middle) and *M. smegmatis* MC²155 (lowest). Locally collinear blocks (LCBs) show the regions of homology among all the three strains. LCB connection lines trace and connect each orthologous LCB thorough all the three genomes. The inverted regions in Mtb H37Rv are clearly shown as blocks below the central line of the genome. The three genome files used for this analysis were downloaded from NCBI GenBank (full) format.

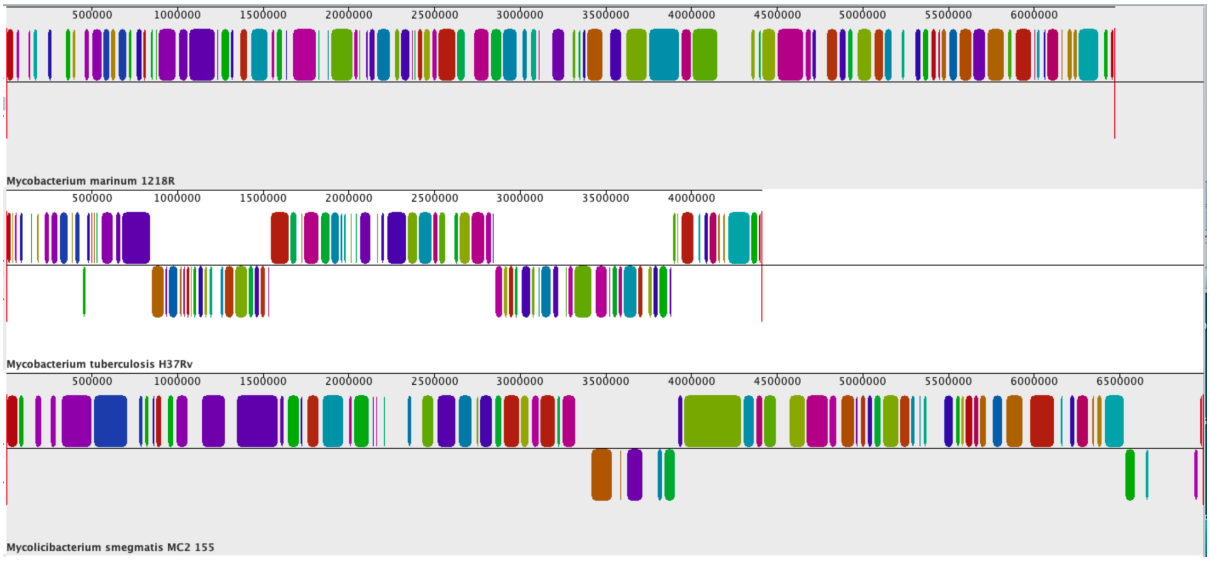


Figure 3.6 Results of the alignment (Figure 3.5) shown without LCB connection lines and with solid LCBs.

3.3.1.8 Roary analysis

It was observed in the ACT (Figure 3.4 A & B) that the comparison view panel showed more matches of *M. tuberculosis* H37Rv with *M. marinum* 1218R than with M. smegmatis MC²155. It was interesting to find the core genes and percentage of sequence identity of all the three genomes. To investigate this, Roary pangenome analysis was used (Page et al., 2015). Roary is a useful tool to investigate large-scale pangenome analysis which identifies core and accessory genome. The core genome is the set of all the genes shared among all the members under study whereas the accessory genome is the set of genes present in only one or more strains under study. To run the analysis FASTA genome files of 1218R, H37Rv and MC²155 downloaded from NCBI were reannotated using PROKKA (Rapid Prokaryotic Genome Annotation). GFF3 annotation files generated were used to run the Roary analysis. Roary translates all the coding sequences into protein sequences and generates gene clusters. The program then looks for the homologous genes using blastp. And finally, it creates a file of gene presence/absence

Roary generated gene presence/absence matrix in the csv file format which was used further to create the Roary matrix (Figure 3.7). The pairwise comparison of *M. marinum* 1218R (Mm 1218R), *M. tuberculosis* H37Rv (Mtb H37Rv) and M. smegmatis MC²155 (Msmeg MC²155) showed that Mm 1218R and Mtb H37Rv had the highest percentage of homologous genes. *M marinum* genome was found to be more closely related to *M. tuberculosis* than either are to *M. smegmatis*.

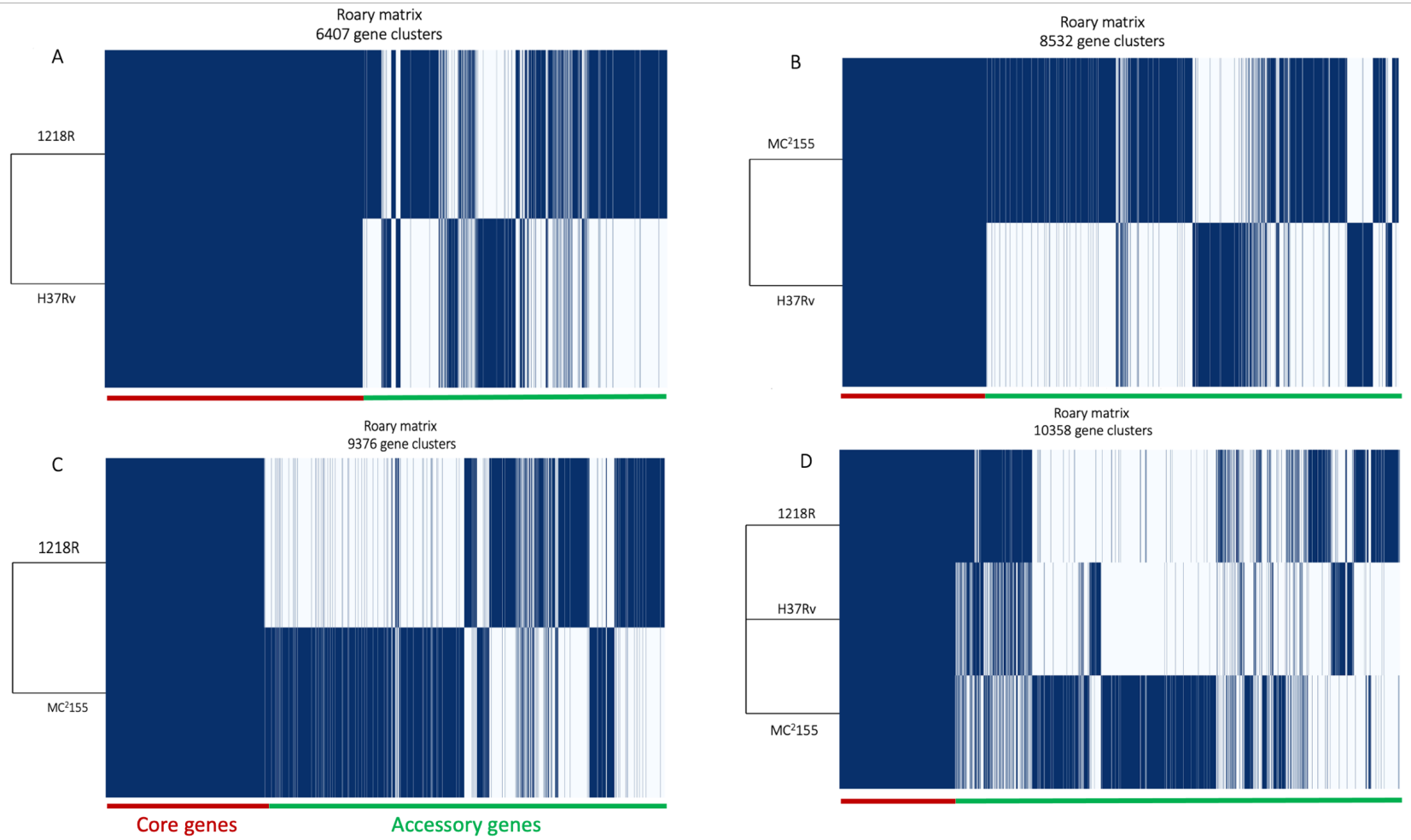


Figure 3.7 Pangenome analyses of *M. marinum* 1218R, *M. tuberculosis* H37Rv, and *M. smegmatis* MC²155 using Roary 60% cut off.
Graphical representation (Roary matrix – gene clusters) of core and accessory genes created using roary_plots.py script. A) 1218R vs H37Rv, B) MC²155 and H37Rv, C) 1218R v MC²155 and D) 1218R, H37Rv and MC²155.

Strains	Total genes	Core genes	Accessory genes	% Match
1218R , H37Rv	6407	2938	3469	45.85
MC²155, H37Rv	8532	2211	6321	25.91
1218R, MC²155	9376	2658	6718	28.34
1218R , H37Rv, MC²155	10276	2140	8135	20.82

Table 3.1 Table showing the core and accessory genes obtained by pangenome analyses of *M. marinum* 1218R, *M. tuberculosis* H37Rv, and *M. smegmatis* MC²155.
Mm 1218R and Mtb H37Rv genomes showed to have the highest percentage of core genes i.e., 45.85%.

Venn diagram of core, accessory and unique genes in 1218R, H37Rv, and MC2155 according to Roary Analysis output.

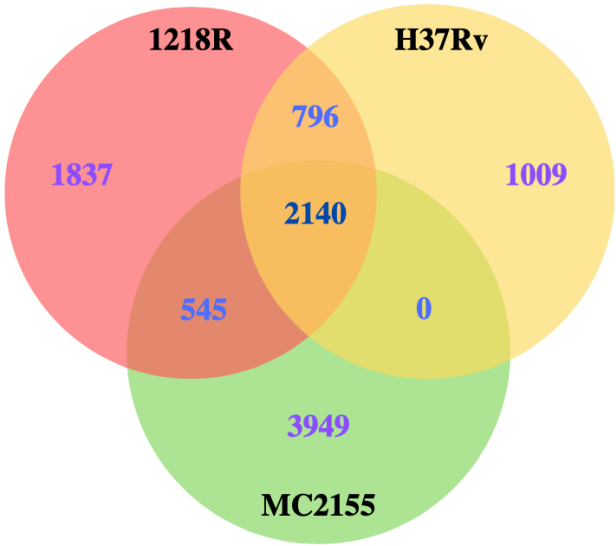


Figure 3.8 Venn diagram showing the results of genome comparison among *M. marinum* 1218R, *M. tuberculosis* H37Rv, and *M. smegmatis* MC²155 using Roary analysis.
2140 core genes were found to be common among all the 3 strains. Mm 1218R and Mtb H37Rv showed 796 genes in common for which no homologue was found in MC²155 whereas 545 genes found in common between Mm 1218R and M. smegmatis MC²155 did not have a homologue in H37Rv. MC²155 and H37Rv did not show any such common gene for which there was no homologue in 1218R. Highest number of unique genes were found in M. smeg MC²155 genome. Overall, the Venn diagram shows that Mm 1218R and Mtb H37Rv genomes share the highest number of homologous genes compared to the other two comparisons i.e., 1218R V MC²155 and H37Rv V MC²155.

3.3.2 Phylogenetic analysis

Phylogenetic analysis provides the basic understanding of evolutionary relationships among species. The phylogenetic trees provide the historical explanation of the similarities and differences among species. The phylogenetic relationship among different mycobacteria was inferred based on the nucleotide sequences of 16S rRNA gene and hsp60 chaperonin (*cpn*) genes. For more precise results the full coding sequence of the 16S rRNA was used which provides more accurate results. For this purpose, 50 species of mycobacteria including *M. marinum*, *M. tuberculosis*, *M. smegmatis* and *M. bovis*, 12 other species of actinobacteria such as *Streptomyces griseus*, *Corynebacterium diphtheriae* and *Mantella aurantiaca* were included in the analysis. Three species of non-actinobacteria, *Escherichia coli*, *Bacillus subtilis* and *Staphylococcus aureus* were used as outgroup organisms. All the sequences were manually collected from the whole genomes of each species which were downloaded from NCBI genome database. The full name of the strains and NCBI accession numbers of each species used in the analysis are given in the table 3.2. The nucleotide sequences of 16S rRNA and all the *cpn60* genes present in a species were extracted from the genomes. The analysis was done using MEGA11 software. Sequences were first aligned using default parameters of muscle alignment and phylogenetic trees were constructed using maximum-likelihood statistical method with 1000 bootstrap iterations.

Figure 3.9 shows the phylogenetic tree based on full-length 16S rRNA gene sequence of all the 65 strains used in the analysis. *M. marinum* and *M. tuberculosis* are marked with green and red dots. The tree shows the evolution of *M. marinum* and *M. tuberculosis* from a more recent common ancestor however, the closest relative of *M. marinum* is *M. ulcerans* (both cause skin disease in humans). *M. tuberculosis* complex which includes *M. tuberculosis*, *M. bovis* and *Mycobacterium africanum* cluster is clearly distant from the fast growers such as *M. smegmatis*. The tree is rooted on *E. coli* (the Proteobacteria), the two Firmicutes *B. subtilis* and *S. aureus* are closest to the *E. coli* in the phylogeny. The tree shows that *M. marinum* common ancestor with *M. tuberculosis* is more recent compared to the other well-known mycobacterial pathogens such as *M. leprae* and *M. avium*. All other actinobacteria species make separate clusters than mycobacteria closer to the tree root indicating the evolution from more distant common ancestors. Most of the fast-growing mycobacteria are closer to the tree root

whereas most of the slow growing mycobacteria are distant from the root showing the evolution from more recent common ancestors.

Figure 3.10 shows the phylogenetic tree based on all the *cpn60* genes present in the 65 species used in the analysis (Table 3.2). As described in chapter 1 (1.4.2), most of the mycobacterial and actinobacterial species have two chaperonin genes (called *cpn60.1* and *cpn60.2*), some have three chaperonin genes (*cpn60.1, cpn60.2* and *cpn60.3*). The two species of mycobacteria *M. smegmatis* and *Mycobacterium goodii* possess three chaperonin genes whereas all other actinobacterial species (including mycobacteria) included in this analysis have two chaperonin genes. The proteobacteria and firmicutes included in this analysis have only one copy of the chaperonin gene (*groEL*).

All the three chaperonins fell into three distinct clades, with the Cpn60.2 clade making an outgroup in the tree (Figure 3.10). The tree is rooted on *E. coli groEL* gene. The two larger clades of Cpn60.1 and Cpn60.2 show that a gene duplication event is likely to have taken place in the common ancestor which was followed by the sequence and function divergence which was later kept conserved following the speciation events. In terms of closeness, in both the Cpn60.1 and Cpn60.2 clades, *M. marinum* makes the closest group with *Mycobacterium shottsii* and both share a recent common ancestor with *M. ulcerans.* However, the group is not far away from *M. tuberculosis* complex. As observed in rRNA tree, in Cpn60.1 clade, *M. marinum* shares a more recent common ancestor with *M. tuberculosis* than other well-known mycobacterial pathogens such as *M. avium*, *M. leprae. M. smegmatis* is further away from *M. tuberculosis* in both Cpn60.1 and Cpn60.2 clades. The phylogenetic trees based on Cpn60.1 and Cpn60.2 sequences alone (Figure 3.11 & Figure 3.12) showed results similar to that of rRNA which indicates that there is no obvious evidence for any horizontal gene transfer.

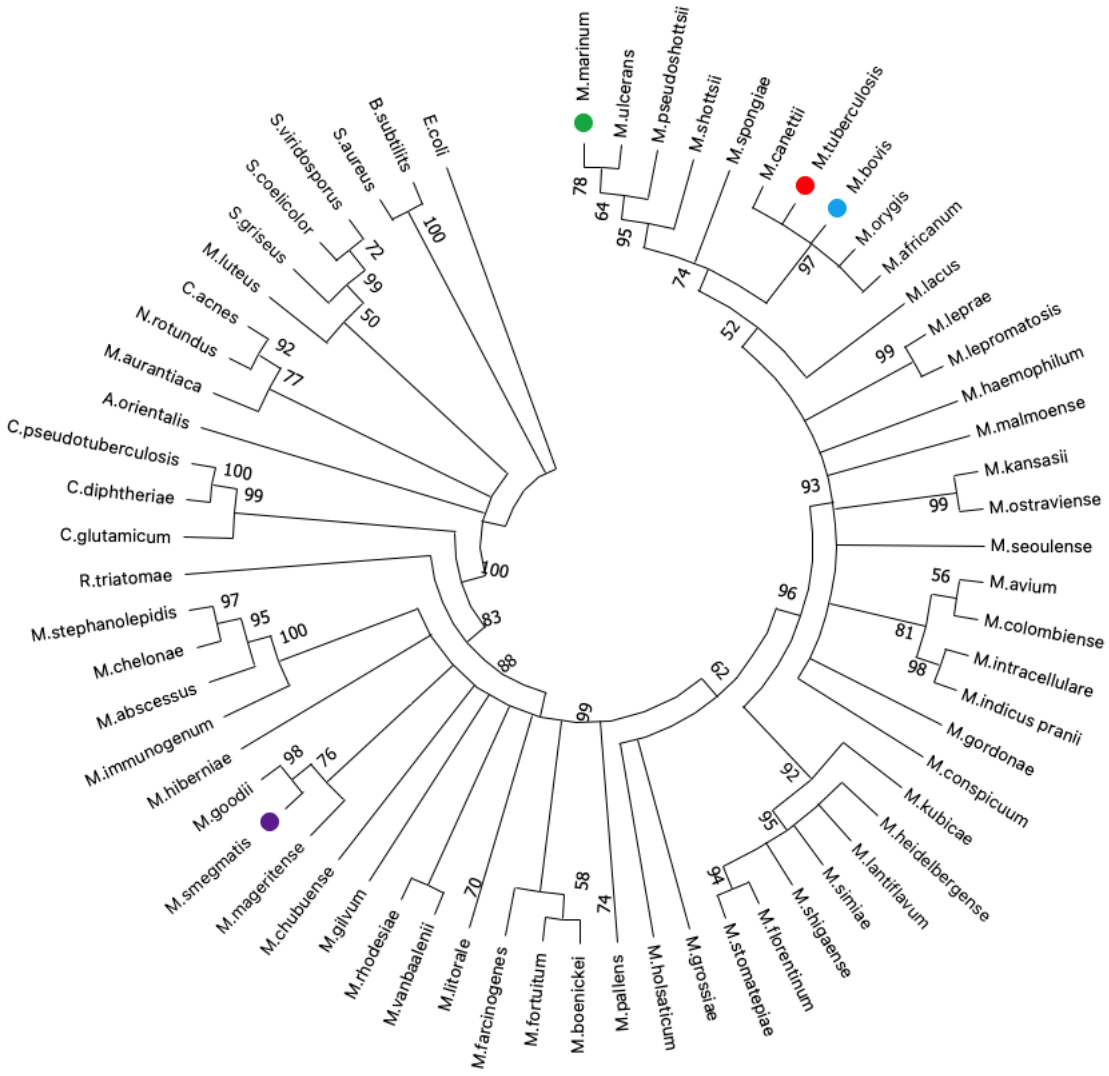


Figure 3.9 The Phylogenetic tree based on 16S rRNA coding gene sequence.
The tree was constructed using maximum-likelihood method (Tamura-Nei model). The numbers show the bootstrap value as a percentage of 1000 replications.

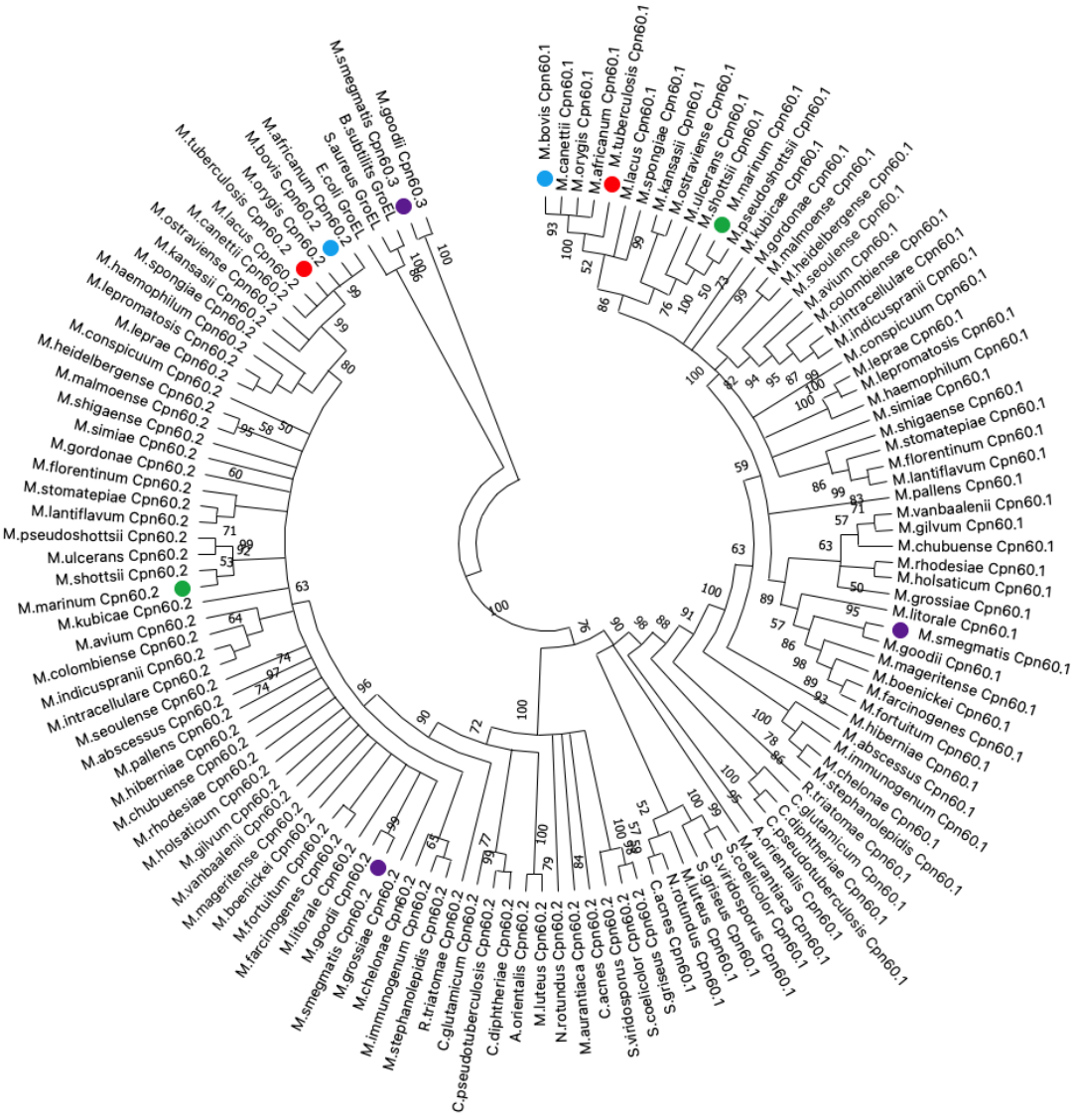


Figure 3.10 Phylogenetic tree based on amino acid sequence of chaperonins.
The Phylogenetic tree based on amino acid sequence of chaperonins present in 50 species of mycobacteria, 12 other species of actinobacteria, two firmicutes (*B. subtilis* and *S. aureus*) and one proteobacterium (*E. coli*). The tree was constructed using maximum-likelihood method. The numbers show the bootstrap value as a percentage of 1000 replications.

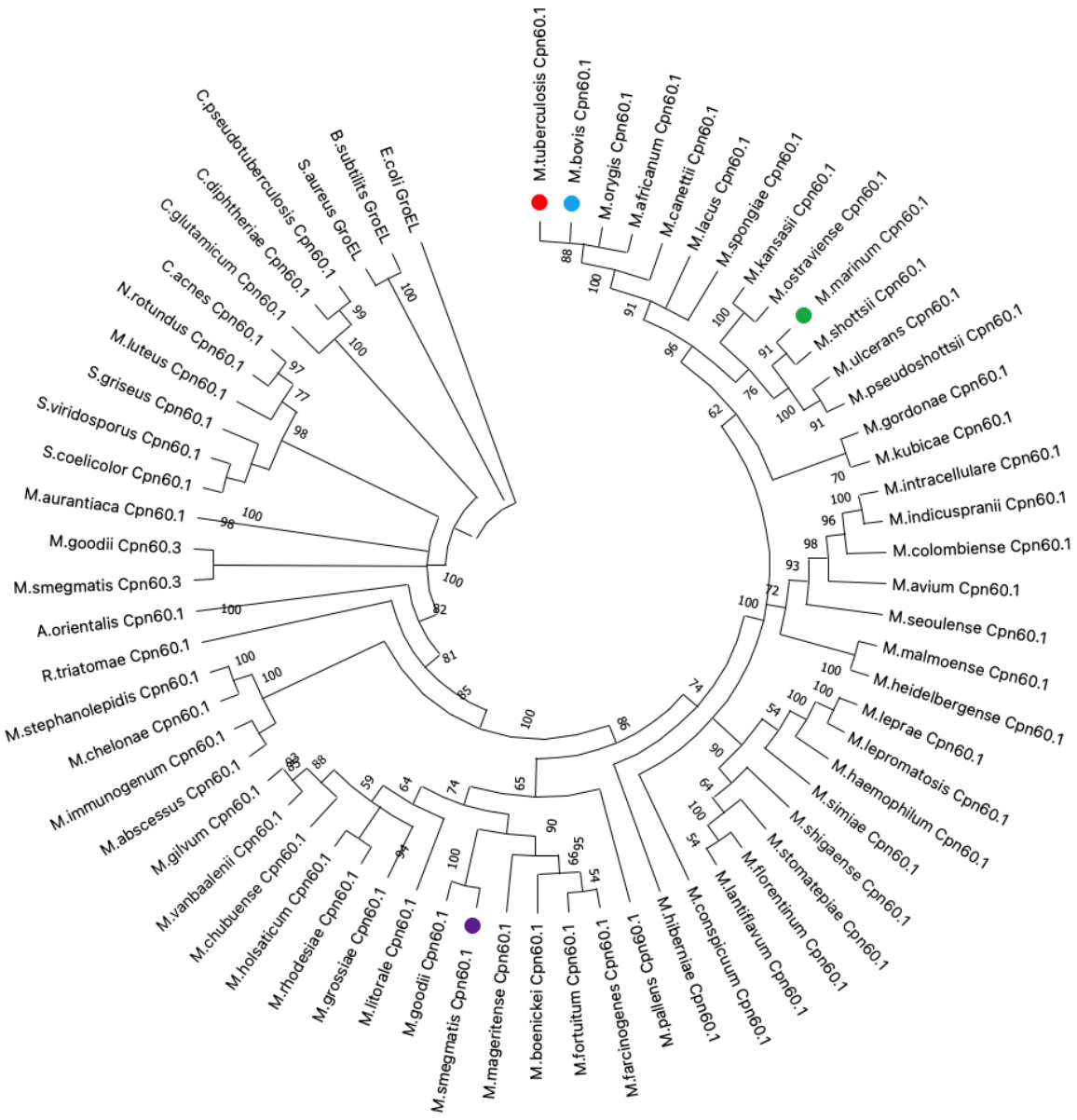


Figure 3.11 Phylogenetic tree based on amino acid sequence of Cpn60.1 and Cpn60.3. Phylogenetic tree based on amino acid sequence of Cpn60.1 and Cpn60.3 present in 50 species of mycobacteria 12 other species of actinobacteria, two firmicutes (*B. subtilis* and *S. aureus*) and one proteobacterium (*E. coli*).

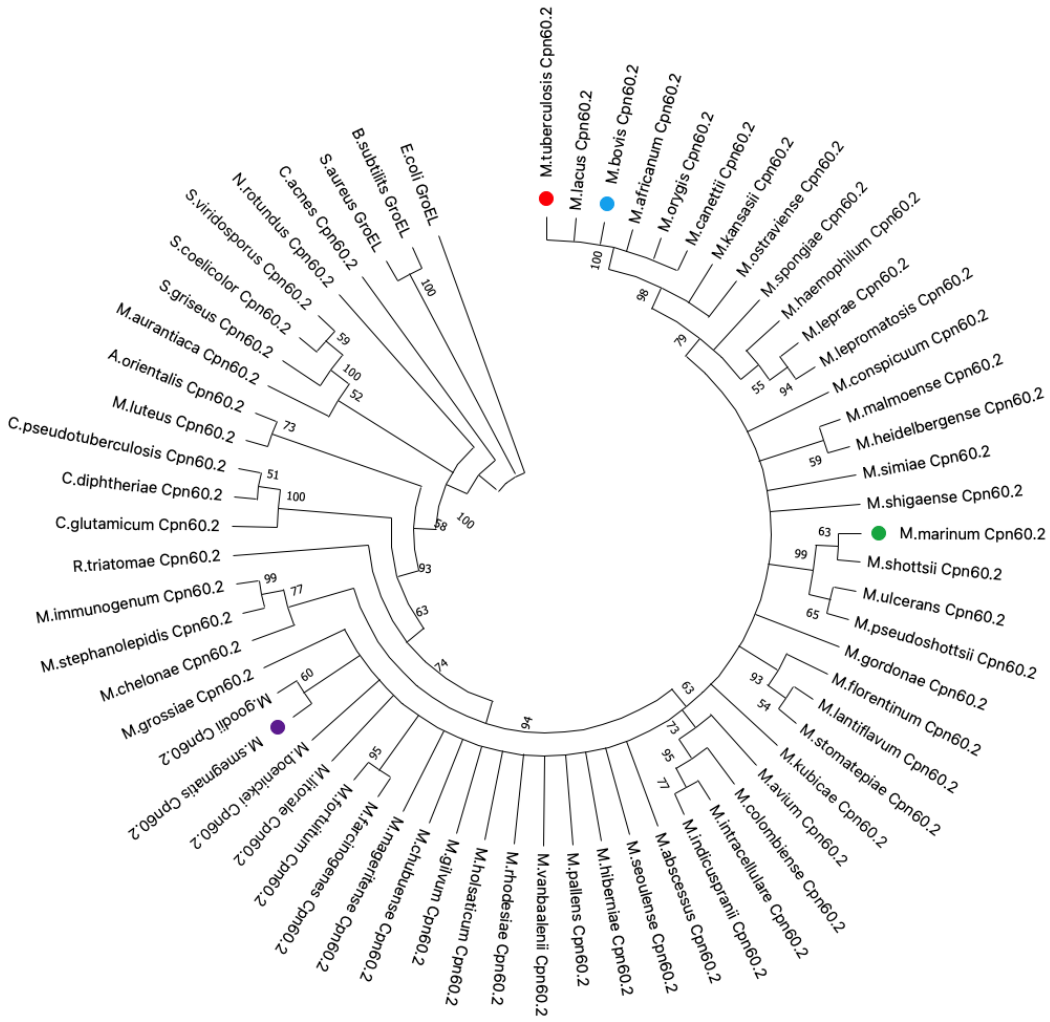


Figure 3.12 The Phylogenetic tree based on amino acid sequence of Cpn60.2.
The Phylogenetic tree based on amino acid sequence of Cpn60.2 present in 50 species of mycobacteria 12 other species of actinobacteria, two firmicutes (*B. subtilis* and *S. aureus*) and one proteobacterium (*E. coli*).

3.3.3 Multiple sequence alignment of chaperonins (Cpn60)

The amino acid sequences of all the chaperonins of 10 well-known species of mycobacteria were aligned to find the sequence diversity and similarity. The ten species included in this analysis are *M. marinum, M. tuberculosis, M. bovis, M. smegmatis, Mycobacterium leprae, Mycobacterium ulcerans, Mycobacterium haemophilum, Mycobacterium abscessus, Mycobacterium lepromatosis* and *M. avium*. *Escherichia coli* GroEL was included in the analysis for comparison. The multiple sequence alignment of all the protein sequences was performed in DNAstar Megalign pro software

(https://www.dnastar.com/software/lasergene/megalign-pro/) using default parameters of MUSCLE algorithm.

The pairwise amino acids sequence identity (%) and similarity (%) for the entire length of the cpn60 homologues was also calculated using the same software and is shown in Figure 3.10. GroEL showed similarity between 73.78 to 74.53% with all the Cpn60.2 proteins whereas with Cpn60.1 proteins it showed a similarity range between 68.73 to 72.10%. The data showed that GroEL has lowest similarity with Cpn60.3 (of *M. smegmatis*) which is 65.54%.

The Cpn60.1 and Cpn60.2 of *M. marinum* and *M. tuberculosis* were found to be 76.03% and 75.28% similar (respectively) with each other. The comparison of all the Cpn60.1 and Cpn60.2 proteins showed the lowest similarity of 73.78%.

The lowest % similarity calculated among all the protein sequences under study was 65.36 that is for the pair Cpn60.3 and Cpn60.1 (of *Mycobacterium lepromatosis*), while the minimum % identity was 48.31 again for Cpn60.3 in comparison with GroEL. Overall Cpn60.3 is shown to be least similar with all the proteins under study (Figure 3.13).

To look at the regions of diversity, the alignment sequence of all the proteins was observed. The highest diversity was observed in the C-terminal of the proteins. Interestingly, the C-terminal of all the Cpn60.2 was found to be rich in glycine and methionine (GGM) amino acids like the C-terminal of GroEL whereas, the C-terminal of all the Cpn60.1 was found to be rich in histidine (H) residues. The tail of Cpn60.2 of *M. tuberculosis* and *M. bovis* was GGDMGGM whereas the tail of all other Cpn60.2 proteins was GGMGGM (no aspartic acid (D) residue in the GGM repeat). Four GGM repeats were found in *E. coli* GroEL C-terminal tail (GGMGGMGGMGGM). Cpn60.3 has diverse C-terminal which does not match with any the chaperonins under study, however it has two glycine and one methionine in the C-terminal tail (Figure 3.14).

		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
A	M.marinum_Cpn60.2		96.63	96.63	97.38	96.82	99.81	98.13	97.94	97.00	99.25	73.97	76.03	75.47	75.47	76.59	76.03	76.03	76.40	75.66	76.03	76.40	68.91
B	M.tuberculosis_Cpn60.2	95.32		100.00	96.07	96.82	96.44	97.94	96.44	96.63	96.82	74.53	76.22	75.28	75.28	76.40	76.03	76.22	76.22	76.22	75.84	76.59	69.85
C	M.bovis_Cpn60.2	95.32	100.00		96.07	96.82	96.44	97.94	96.44	96.63	96.82	74.53	76.22	75.28	75.28	76.40	76.03	76.22	76.22	76.22	75.84	76.59	69.85
D	M.smegmatis_Cpn60.2	95.69	94.38	94.38		96.25	97.19	97.57	98.31	96.44	97.38	74.34	77.53	76.03	76.03	77.53	76.59	77.53	76.97	76.97	76.22	77.34	69.48
E	M.leprae_Cpn60.2	94.57	95.51	95.51	93.63		97.00	98.50	96.82	98.69	97.38	73.78	76.03	74.34	74.34	76.40	75.84	76.03	75.66	76.22	75.47	76.22	69.29
F	M.ulcerans_Cpn60.2	99.63	94.94	94.94	95.32	94.57		97.94	97.75	97.19	99.06	73.97	75.84	75.28	75.28	76.59	75.84	75.84	76.22	75.47	75.84	76.22	69.10
G	M.haemophilum_Cpn60.2	96.44	96.82	96.82	94.76	97.00	96.07		97.94	98.69	98.31	74.34	76.40	75.47	75.47	76.78	76.59	76.40	76.78	76.59	76.40	77.15	69.85
H	M.abscessus_Cpn60.2	96.44	94.76	94.76	95.69	94.19	96.25	94.94		96.82	98.13	73.97	76.59	75.47	75.47	77.34	76.59	76.59	76.97	76.40	76.22	76.97	69.29
I	M.lepromatosis_Cpn60.2	94.57	95.32	95.32	94.01	97.57	94.57	97.75	94.19		97.19	73.97	76.22	75.09	75.09	76.59	76.40	76.22	76.59	76.59	76.22	76.97	69.66
J	M.avium_Cpn60.2	97.57	95.13	95.13	94.94	94.38	97.19	95.69	95.69	94.57		73.97	76.03	75.09	75.09	76.59	75.84	76.03	76.22	75.84	75.66	76.22	69.29
K	E.coli_GroEL	59.36	59.74	59.74	59.55	59.18	59.18	59.74	59.36	59.55	59.74		71.16	68.73	68.73	72.10	70.97	71.16	71.16	71.16	70.04	71.35	65.54
L	M.marinum_Cpn60.1	60.67	61.05	61.05	61.61	60.49	60.30	61.24	61.05	60.86	60.86	53.75		94.01	94.01	92.51	93.07	100.00	94.76	90.45	91.95	93.82	66.85
M	M.tuberculosis_Cpn60.1	61.24	61.42	61.42	61.99	60.30	60.86	61.24	61.24	60.86	60.86	52.43	87.08		100.00	91.39	92.13	94.01	94.19	88.20	91.20	92.32	64.98
N	M.bovis_Cpn60.1	61.24	61.42	61.42	61.99	60.30	60.86	61.24	61.24	60.86	60.86	52.43	87.08	100.00		91.39	92.13	94.01	94.19	88.20	91.20	92.32	64.98
O	M.smegmatis_Cpn60.1	62.17	62.92	62.92	63.48	62.17	61.99	62.73	62.92	62.36	62.55	55.06	82.21	81.84	81.84		91.57	92.51	91.76	91.76	90.07	92.32	66.67
P	M.leprae_Cpn60.1	60.30	60.67	60.67	61.24	60.11	59.93	60.86	60.67	60.30	60.11	52.25	84.08	83.33	83.33	80.52		93.07	96.82	89.33	96.82	93.07	65.92
Q	M.ulcerans_Cpn60.1	60.30	60.49	60.49	61.05	59.93	59.93	60.67	60.67	60.30	60.49	53.37	99.44	86.52	86.52	81.65	83.52		94.76	90.45	91.95	93.82	66.85
R	M.haemophilum_Cpn60.1	61.61	61.80	61.80	62.73	60.86	61.24	61.99	61.99	61.61	61.24	52.81	85.96	86.14	86.14	81.09	94.01	85.39		89.14	95.88	94.19	65.92
S	M.abscessus_Cpn60.1	61.99	62.36	62.36	62.73	61.80	61.61	61.99	62.55	61.99	62.17	53.18	79.21	78.09	78.09	83.52	78.84	78.65	79.40		87.83	90.07	66.48
T	M.lepromatosis_Cpn60.1	59.74	60.11	60.11	60.49	59.74	59.36	60.30	59.93	59.93	59.55	51.69	82.77	82.77	82.77	78.84	93.63	82.21	92.70	77.15		91.57	65.36
U	M.avium_Cpn60.1	62.73	62.73	62.73	62.73	61.61	62.36	62.92	62.36	62.36	62.36	53.37	85.02	84.08	84.08	81.84	85.21	84.46	87.27	81.27	83.90		66.85
V	M.smegmatis_Cpn60.3	50.94	51.87	51.87	51.31	50.75	50.75	50.94	50.94	50.75	51.50	48.31	47.75	46.82	46.82	49.06	47.57	47.19	48.50	49.63	47.38	48.50	

Figure 3.13 Pairwise amino acid identity (%) and similarity (%) for homologous proteins.

Pairwise amino acid identity (%) and similarity (%) for the full length of Cpn60 homologous protein sequences from 10 species of mycobacteria and *E. coli* GroEL. The calculations were done using DNAstar Megalign pro software. Red colour boxes indicate % identity whereas turquoise colour boxes show % similarity. The intensity of the colour is directly proportional to the %age of the identity or similarity.

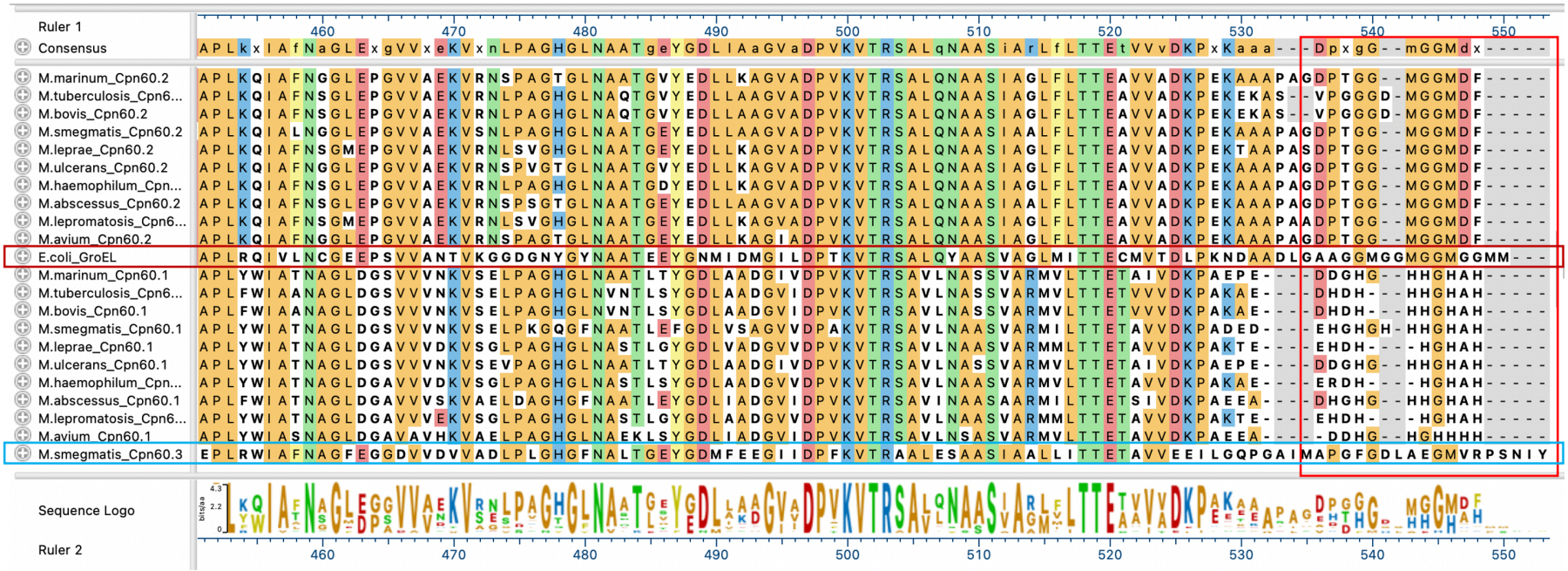


Figure 3.14 Multiple sequence alignment of Cpn60 homologues of ten mycobacterial species and *E. coli* GroEL.
Only C-terminus residues are shown. The box at the tail highlights the G-M rich C-terminal of Cpn60.2 and GroEL proteins and histidine rich C-terminal of Cpn60.1 proteins. GroEL in the middle of all the Cpn60.2 and Cpn60.1 is also boxed. At the bottom of the alignment is Cpn60.3.

The alignment shows a clear diversity of C-termini among Cpn60.1 and Cpn60.2 chaperonins. To look if this diversity is unified across other mycobacterial and actinobacterial species, the Cpn60.1 and Cpn60.2 amino acid sequences of 63 actinobacterial species (50 mycobacterial and 12 other actinobacterial) were used in the analysis. The strains used are given in the Table 3.2. The multiple sequence alignment of all the protein sequences was performed using the same software and algorithm as above. The alignment of Cpn60.2 residues is shown in Figure 3.15 and that of Cpn60.1 is displayed in Figure 3.16.

The pattern of GGDMGGM found in *M. tuberculosis* and *M. bovis* was found in six other mycobacterial species whereas all other mycobacteria have GGMGGM pattern. If the presence of aspartic acid in the C-terminal tail of these species has any significance is not known yet. Moreover, the GroEL has 4 GGM repeats whereas no other actinobacteria was found to have more than two GGM repeats. The lowest GGM residues were found in the C-terminal of three species of Streptomyces and three species of Corynebacterium (Figure 3. 15). The C-terminal tail of all the Cpn60.1 proteins under study were found to be histidine rich. No highly conserved histidine pattern was observed in the alignment. All the Cpn60.1 proteins were observed to have more histidine residues at the tail rather than glycine residues.

Overall, we can summarize that Cpn60.2 C-terminal tail was found to be rich in glycine methionine residues like GroEL C-terminal whereas the C-terminal tail of Cpn60.1 was found to be histidine rich.

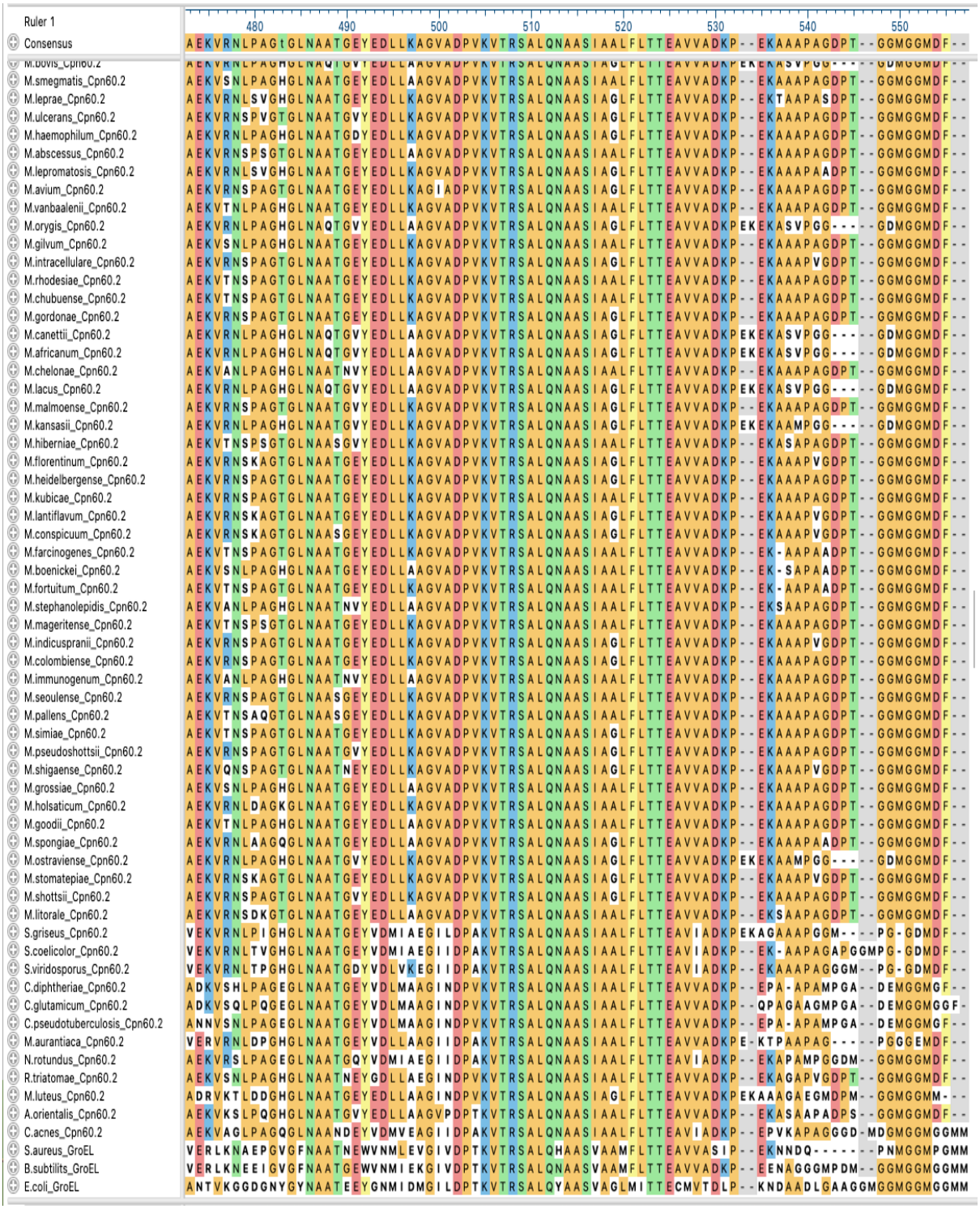


Figure 3.15 Multiple sequence alignment of Cpn60.2 homologues.
Multiple sequence alignment of Cpn60.2 homologues of 50 mycobacterial species, 12 other actinobacterial species and three GroEL proteins from *S. aureus*, *B. subtilis* and *E. coli*. Only C-terminus residues are shown. Consecutive yellow colour rows at the end of the C-terminal highlight glycine methionine residues.

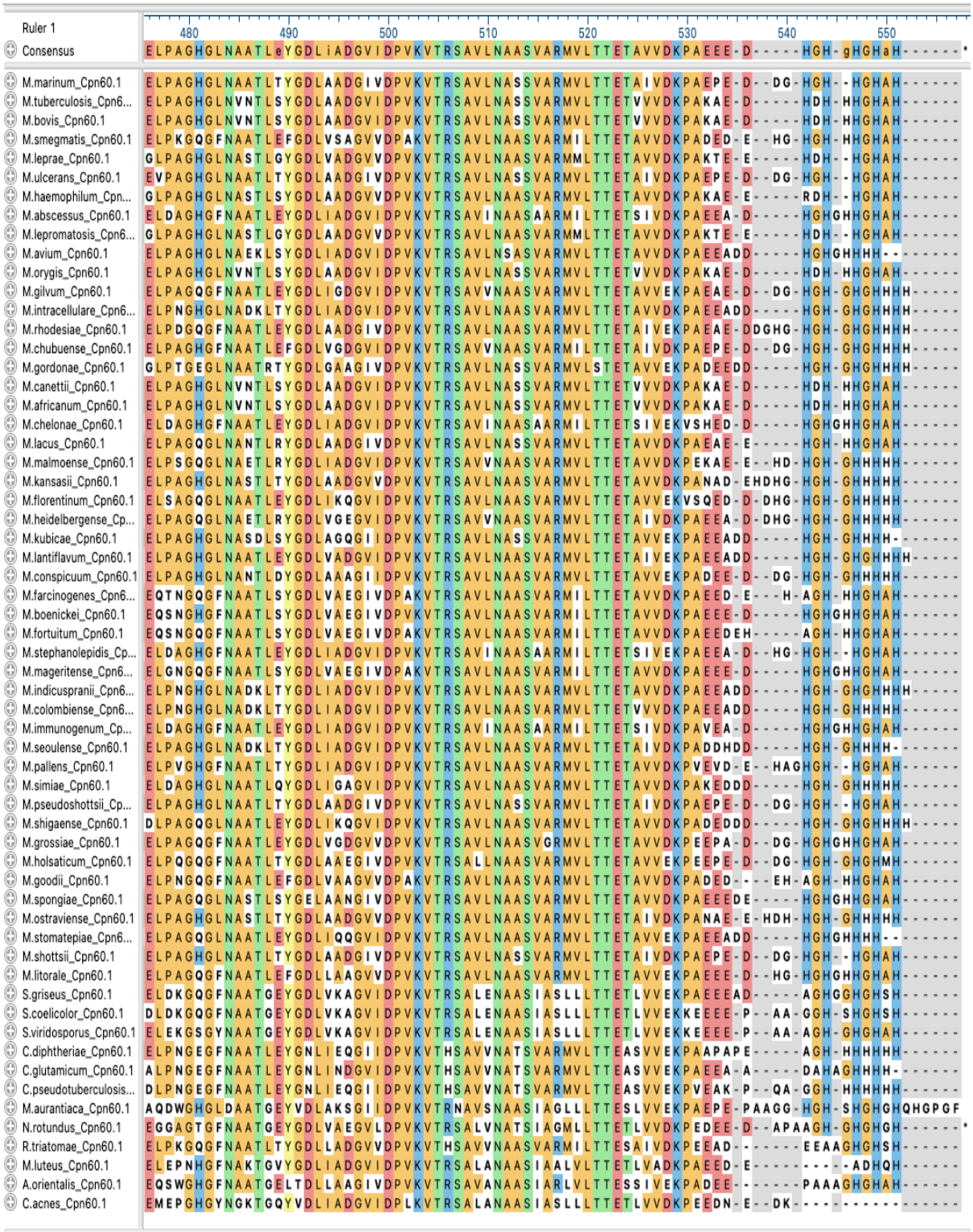


Figure 3.16 Multiple sequence alignment of Cpn60.1 homologues.
Multiple sequence alignment of Cpn60.1 homologues of 48 mycobacterial species and 12 other actinobacterial species. Only C-terminus residues are shown. Blue colour rows at the end of the C-terminal highlight histidine residues.

3.4 Discussion

The genome comparison of *M. tuberculosis* and *M. marinum* has shown that *M. tuberculosis* has reduced its genome to specialise for host specificity and pathogenesis. Whereas *M. marinum* has kept the large genome to survive in a variety of host environments. For example, *M. tuberculosis* genome possess only 13 pseudogenes whereas, *M. marinum* genome has 192 pseudogenes. Both organisms have a large number of genes whose function is still not known such as PE/PPE family genes and genes encoding hypothetical proteins. More detailed study of the genes of unknown functions would be advantageous in understanding the nature of this incredibly successful pathogen.

The phylogenetic analysis based on rRNA and chaperonin genes of 50 species of mycobacteria and 12 other species of actinobacteria showed that *M. marinum* share a more recent common ancestor with *M. tuberculosis* compared to many other well-known mycobacterial pathogens and model organisms such as *M. leprae, M. avium, and M. smegmatis*. According to the Roary Pangenome analysis *M. marinum* share more core genes (~50% of the total genes in the Roary matrix of *M. marinum* and *M. tuberculosis*) with *M. tuberculosis* compared to *M. smegmatis*. We therefore concluded that *M. marinum* is a closer genetic relative of *M. tuberculosis* than *M. smegmatis*.

The two chaperonin proteins in all the mycobacteria possess at least 60% amino acid identity with higher divergence and dissimilarity at the C-terminal. Diversity in the C-terminals of *Cpn60.1* and *Cpn60.2* is interesting to investigate further. The glycine rich C-terminal of Cpn60.2 matches more closely with GroEL whereas the c-terminal of Cpn60.1 is histidine rich. The phylogenetic tress of 62 actinobacteria based on all the chaperonin genes present in them suggested that *cpn60.1* and *cpn60.2* might have arisen from a single ancestor gene duplication event and no obvious evidence of horizontal gene transfer was observed.

The following table contains the names and NCBI accession numbers of the species used for phylogenetic and sequence analyses.

No.	Phylum	Organism	NCBI accession no.	G+C content	No. of chaperonin genes	Containment level	Growth rate
1	Actinobacteria	*Mycobacterium marinum* 1218R	NZ_CP025779.1	65.7	2	2	Slow
2	Actinobacteria	*Mycobacterium tuberculosis* H37Rv	AL123456.3	65.61	2	3	Slow
3	Actinobacteria	*Mycobacterium bovis* BCG	CP002095.1	65.66	2	2	Slow
4	Actinobacteria	*Mycobacterium smegmatis* MC2 155	CP000480.1	67.4	3	1	Fast
5	Actinobacteria	*Mycobacterium leprae* TN	AL450380.1	57.8	2	3	Slow
6	Actinobacteria	*Mycobacterium ulcerans* JKD8049	CP085200.1	65.49	2	3	Slow
7	Actinobacteria	*Mycobacterium haemophilum* DSM 44634 ATCC 29548	NZ_CP011883.2	63.95	2	2	Slow
8	Actinobacteria	*Mycobacteroides abscessus* GZ002	NZ_CP034181.1	64.13	2	2	Fast
9	Actinobacteria	*Mycobacterium lepromatosis* FJ924	CP083405.1	57.89	2	2	Slow
10	Actinobacteria	*Mycobacterium avium* K10	AE016958.1	69.3	2	2	Slow
11	Actinobacteria	*Mycobacterium vanbaalenii* PYR-1	CP000511.1	67.79	2	1	Fast
12	Actinobacteria	*Mycobacterium orygis* 51145	CP063804.1	65.62	2	2	Slow
13	Actinobacteria	*Mycobacterium gilvum* Spyr1	CP002385.1	67.86	2	1	Fast
14	Actinobacteria	*Mycobacterium intracellulare* MOTT-64	CP003324.1	68.07	2	2	Slow
15	Actinobacteria	*Mycobacterium rhodesiae* NBB3	CP003169.1	65.49	2		Fast
16	Actinobacteria	*Mycobacterium chubuense* NBB4	CP003053.1	68.69	2		Fast
17	Actinobacteria	*Mycobacterium gordonae* X7091	NZ_CP070973.1	66.73	2	2	Slow
18	Actinobacteria	*Mycobacterium canettii* CIPT 140010059	HE572590.1	65.62	2		Slow

19	Actinobacteria	*Mycobacterium africanum* GM041182	FR878060.1	65.59	2	3	Slow
20	Actinobacteria	*Mycobacterium chelonae* DSM 43804	NZ_CP034383.1	63.92	2	2	Fast
21	Actinobacteria	*Mycobacterium lacus* JCM 15657	NZ_AP022581.1	66.95	2		Slow
22	Actinobacteria	*Mycobacterium malmoense* ATCC 29571	NZ_CP080999.1	66.99	2	3	Slow
23	Actinobacteria	*Mycobacterium kansasii* ATCC 12478	CP006835.1	66.24	2	2	Slow
24	Actinobacteria	*Mycolicibacter hiberniae* JCM 13571	AP022609.1	68.51	2		Slow
25	Actinobacteria	*Mycobacterium florentinum* JCM 14740	NZ_AP022576.1	66.38	2		Slow
26	Actinobacteria	*Mycobacterium heidelbergense* JCM 14842	AP022615.1	67.94	2		Slow
27	Actinobacteria	*Mycobacterium kubicae* JCM 13573	NZ_CP065047.1	66.02	2		Slow
28	Actinobacteria	*Mycobacterium lentiflavum* ATCC 51985	CP092423.1	66.04	2	2	Slow
29	Actinobacteria	*Mycobacterium conspicuum* JCM 14738	AP022613.1	67.39	2		Slow
30	Actinobacteria	*Mycolicibacterium farcinogenes* BKK/CU-MFGLA-001	NZ_CP080510.1	66.4	2		Slow (similar to Fast)
31	Actinobacteria	*Mycolicibacterium boenickei* PDNC014	NZ_CP070348.1	66.46	2		Fast
32	Actinobacteria	*Mycolicibacterium fortuitum* JCM6387	NZ_AP025518.1	66.19	2	2	Fast
33	Actinobacteria	*Mycobacterium stephanolepidis*	AP018165.1	63.95	2		
34	Actinobacteria	*Mycolicibacterium mageritense* JCM 12375	NZ_AP022567.1	66.92	2		
35	Actinobacteria	*Mycobacterium indicus pranii* MTCC 9506	CP002275.1	68.03	2		Slow
36	Actinobacteria	*Mycobacterium colombiense* CECT 3035	NZ_CP020821	68.09	2		
37	Actinobacteria	*Mycobacterium immunogenum* CCUG 47286	CP011530.1	64.27	2		Fast

38	Actinobacteria	*Mycobacterium seoulense* JCM 16018	NZ_AP022582.1	68.29	2		
39	Actinobacteria	*Mycobacterium pallens* JCM 16370	CP080333.1	66.44	2		
40	Actinobacteria	*Mycobacterium simiae* JCM 12377	AP022568.1	66.22	2	2	Slow
41	Actinobacteria	*Mycobacterium pseudoshottsii* JCM 15466	AP018410.1	65.64	2		Slow
42	Actinobacteria	*Mycobacterium shigaense* JCM 32072T	AP018164.1	67.26	2		
43	Actinobacteria	*Mycobacterium grossiae* DSM 104744	CP043474.1	70.48	2		
44	Actinobacteria	*Mycobacterium holsaticum* JCM 12374	CP080998.1	67.23	2		
45	Actinobacteria	*Mycobacterium goodii* ATCC 700504	CP092364.1	67.05	3		Fast
46	Actinobacteria	*Mycobacterium spongiae* FSD4b-SM	CP046600.1	65.56	2		
47	Actinobacteria	*Mycobacterium ostraviense* FDAARGOS_1613	CP089224.1	66.27	2		
48	Actinobacteria	*Mycobacterium stomatepiae* JCM 17783	AP022587.1	65.96	2		
49	Actinobacteria	*Mycobacterium shottsii* JCM 12657	AP022572.1	65.53	2		Slow
50	Actinobacteria	*Mycolicibacterium litorale* JCM 17423	NZ_AP022586.1	68.92	2		Fast
51	Actinobacteria	*Streptomyces griseus* NBRC 13350	NC_010572.1	72.23	2	1	
52	Actinobacteria	*Streptomyces coelicolor* A3(2)	AL645882.2	72.12	2	1	
53	Actinobacteria	*Streptomyces viridosporus* T7A ATCC 39115	NZ_CP023700.1	72.56	2	1	
54	Actinobacteria	*Corynebacterium diphtheriae* CD1008	NZ_CP040525.1	53.55	2	2	
55	Actinobacteria	*Corynebacterium glutamicum* SCgG2	NC_021352.1	53.93	2	1	
56	Actinobacteria	*Corynebacterium pseudotuberculosis* MEX29	NZ_CP016826.1	52.18	2	2	
57	Actinobacteria	*Micromonospora aurantiaca* ATCC 27029	NC_014391.1	72.82	2		

58	Actinobacteria	*Nocardioides rotundus* MCCC 1A10561	NZ_CP082922.1	71.46	2		
59	Actinobacteria	*Rhodococcus triatomae* DSM 44893	NZ_CP048813.1	68.68	2		
60	Actinobacteria	*Micrococcus luteus* CW.Ay	NZ_CP082331.1	73.06	2	1	
61	Actinobacteria	*Amycolatopsis orientalis* B-37	NZ_CP016174.1	68.83	2		
62	Actinobacteria	*Cutibacterium acnes* HL096PA1	NC_021085.1	60.11	2		
63	Firmicutes	Staphylococcus aureus MRSA252	BX571856.1	32.81	1	2	
64	Firmicutes	*Bacillus subtilis* UD1022	CP011534.1	43.89	1	1	
65	Proteobacteria	*Escherichia coli* K-12 MG1655,	NC_000913.3	50.79	1	2	
66	Plasmid	*Mycobacterium marinum* 1218R plasmid pMMRN	AP018497.1				

Table 3.2 Names of the species used for phylogenetic and sequence analysis.

4 Analysis of promoter activity of regions upstream of *cpn* genes in *M. marinum*

4.1 Introduction

As discussed in chapter 1, all known mycobacterial species possess at least two chaperonin homologues, one of which (*cpn60.1*) is always present downstream of the cochaperonin (*cpn10*) in a single operon, and the other of which (*cpn60.2*) is found elsewhere on the chromosome. Repeated experimental evidence demonstrates that *cpn60.2* is essential in all the cases examined so far, while *cpn60.1* is dispensable. Although experimental data has been reported that Cpn60.1 may have specialized "moonlighting" functions, a complete understanding of its role and function is still not available for any of the mycobacterial species.

Work in our laboratory is utilizing the *M. marinum* – zebrafish model system to improve our understanding of the role of Cpn60.1 protein in this species, in the hope that this will provide us insights into the role of this protein in the human pathogen *Mycobacterium tuberculosis*. One way to gain insights into the function of a given gene or a family of genes is to study their expression and look at the impact of their deletion (in the case of non-essential genes) on the expression levels of other genes. However, the expression of the chaperonin genes in *M. marinum*, the subject of this chapter, has not yet been studied. For this reason, work described in this chapter and the next chapter (chapter 5) used two distinct methods to analyse the expression of the chaperonin genes in *M. marinum*: the construction and analysis of promoter probe plasmids and the use of RNA-seq analysis to investigate the impact of *cpn60.1* deletion on overall gene expression. To put the results of this work in context, the section below will describe the published studies on the expression analysis of chaperonin genes in actinobacteria, especially mycobacterial species under normal and different stress conditions such as heat shock.

In many bacterial species where more than one chaperonin genes are present, the differential expression of these genes is generally observed both under normal and different stress conditions. Particularly, the essential copy of the chaperonin gene is always reported to be expressed and induced at higher levels compared to the dispensable copy (Fischer et al., 1993; Hu et al., 2008; Sato et al., 2008). In *Corynebacterium glutamicum,* both the *cpn10-cpn60.1* operon and the *cpn60.2* operon were found to be heat shock inducible (Barreiro et al., 2004). Later, a comparative proteomics and transcriptome analysis with *C. glutamicum* confirmed the heat shock

inducible nature of both chaperonin genes, with *cpn60.2* expressed slightly more strongly than *cpn60.1* (Barreiro et al., 2005). Another example was a study on *Streptomyces lividans* where *cpn60.2* was reported to be more highly expressed than the other *cpn60* gene. Also, considerable upregulation in the transcriptional levels of both *cpn* operons was observed at high temperature; again, the *cpn60.2* was upregulated more strongly than the other homologue (León et al., 1997).

In actinobacteria, all the cases where two chaperonin genes are present in an organism, *cpn60.1* is always present in a single operon with *cpn10* (Figure 1.17 &Figure 1.18). The question then arises whether *cpn10* and *cpn60.1* are expressed as an operon from a single upstream promoter or additional signals controlling the expression of *cpn60.1* exist?

Earlier studies with *M. tuberculosis* using primer extension analysis showed the presence of the 5′ end of the *cpn60.1* mRNA downstream of *cpn10*, it was therefore proposed that might be *cpn10* and *cpn60.1* are co-transcribed but later undergo a specific cleavage event of the bicistronic mRNA leading to the formation of 5′ end of the *cpn60.1* mRNA. Besides, due to the matches with the consensus sequences for σ^{70} and σ^{32} promoters found in the upstream region of *cpn60.1,* it was also suggested that *cpn60.1* might not be co-transcribed with the upstream *cpn10* gene, and that the transcriptional initiation point for *cpn60.1* could be 28 bps upstream of the start codon of *cpn60.1* (Kong et al., 1993). However, studies with *Corynebacterium glutamicum* using internal probe hybridization to total RNA supported the concept of the bicistronic nature of the *cpn10-cpn60.1* mRNA (Barreiro et al., 2004). In attempt to resolve this issue, the following section will review what is known about expression of the chaperonin genes in mycobacteria in more detail.

4.1.1 Expression pattern of chaperonin genes of *M. tuberculosis*

Microarray studies to examine the mRNA levels of *cpn* genes in *M. tuberculosis* showed nearly 9-fold, 7-fold and 3-fold increase in the mRNA levels of *cpn10*, *cpn60.2* and *cpn60.1* respectively following heat shock at 45°C for 30 minutes. Similarly, under different stress conditions such as hyperosmolarity, oxidative stress, starvation and alkaline pH, the expression levels were increased in the same order with the exception of *cpn60.1* which only showed upregulation in case of heat shock and hyperosmolarity. Despite being in a single operon, the expression profile of *cpn10* gene rarely matched with the expression pattern of *cpn60.1*; rather its expression levels matched with those of *cpn60.2* both under normal and stressed conditions, consistent the fact that the two proteins work together to perform the main chaperone function in the cell (Hu et al., 2008).

In another study (Aravindhan et al., 2009), the promoter activity was studied using promoter probe vectors, constructed by cloning the upstream regions of *cpn* genes of *M. tuberculosis* in a promoter-less *lacZ* plasmid. The β-galactosidase activity assay (which was performed in *M. smegmatis*) showed strong promoter activity in the upstream fragments of *cpn10* and *cpn60.2*, while no promoter activity was observed in the intergenic sequence of *cpn10–cpn60.1*. Primer extension study to find the transcription start site (TSS) in the intergenic region did not provide any evidence of the presence of TSS in this region (Aravindhan et al., 2009), favouring the hypothesis that post-transcriptional cleavage of the bi-cistronic mRNA might be the reason for the differential expression of the *cpn10* and *cpn60.1* genes. The lower levels of *cpn60.1* expression compared to *cpn10* could then be explained by the hypothesis that *cpn60.1* mRNA might degrade rapidly following cleavage.

Aravindhan et al., 2009 also studied the induction levels of upstream regions of chaperonin genes of *M. tuberculosis* in *M. smegmatis* under various stress conditions. The *cpn10* promoter was found to be significantly induced (nearly doubled) at heat shock and oxidative stress but the *cpn60.2* promoter was found to be constitutive (Aravindhan et al., 2009). The induction results in this study differ from the earlier results of Hu et al., 2008 which could partly be explained by the possibility that *M. tuberculosis* promoters might not express or function effectively in *M. smegmatis*.

4.1.2 Expression pattern of chaperonin genes of *M. smegmatis*

Several studies have investigated chaperonin expression and function in *M smegmatis*, with a view to better understanding chaperonin function in mycobacteria in general. Work done by a PhD student in our lab used promoter probes and qRT-PCR to study the expression levels of the *cpn* genes in *M. smegmatis*. Under normal growth conditions, *cpn60.2* was found to be the most highly expressed gene, followed by the co-chaperonin *cpn10*, and then *cpn60.1*, while the expression of *cpn60.3* (a third chaperonin gene in *M. smegmatis*) was not detectable. Compared to the normal growth condition a steady increase of 20 to 200-fold induction of all the genes (except *cpn60.3*) was reported after heat shock at 42°C for 5 to 15 minutes; however, after a longer heat shock of 30 minutes the gene expression was seen to decrease back to close to normal levels (nearly 50-fold of the normal growth) (Rao & Lund, 2010).

The degree of heat shock induction seen in the study of Rao & Lund, 2010 for the *M. smegmatis cpn60.2* gene (≈ 50-fold) was much higher than that reported for *cpn60.2* in *M. tuberculosis* under the same conditions (Hu et al., 2008); however, the measurement in the second case was done using microarray analysis, which has a lower dynamic range than qRT-PCR, so this may be a technical artefact.

In the (Rao & Lund, 2010) study, to find the promoter activity of chaperonin genes, different lengths of the upstream regions of all the *cpn* genes were cloned in pSD5B (a promoter-less lacZ plasmid) and lacZ activity was measured following transformation of the reporter plasmids into *M. smegmatis*. Strong promoter activity for the region upstream of *cpn10* and *cpn60.2* was observed, while no promoter activity was found in the intergenic region of *cpn10* and *cpn60.1*. However, a 5′ mRNA end was detected between *cpn10* and *cpn60.1* using 5' RACE, which could have been generated by post-transcriptional cleavage. These findings are consistent with the work of Kong et al., 1993.

4.1.3 RNAseq and Ribosome profiling data of *cpn* genes of *M. tuberculosis* and *M. smegmatis*

I used the online database "Mycobrowser" (Kapopoulou et al., 2011) to search the information available for the expression levels of *cpn* genes in *M. tuberculosis* and *M. smegmatis*: RNA-seq and ribosome profiling graphs of *cpn* genes expression are based on the data provided by the high-throughput studies by Shell et al., 2015 to measure the transcriptome and translatome profiling in the two mycobacterial species. In this study, RNA-seq and ribosome profiling analysis were used to measure the relative levels of genome wide transcription and translation in *M. smegmatis*. Based on the experimental evidence, Shell et al., 2015 proposed that mRNA processing of a subset of polycistronic mRNAs to produce differential expression, as discussed above for the *cpn10-cpn60.1* operon, is a ubiquitous feature in mycobacteria.

The data supported the observation that, in both mycobacterial species, despite being transcribed as a single mRNA, the levels of mRNA measured in RNAseq for *cpn10* were higher than that of *cpn60.1*. This is consistent with the hypothesis that the bicistronic mRNA produced by the co-transcription of *cpn10* and *cpn60.1* undergoes post-transcriptional cleavage. While ribosome profiling suggests that out of two mRNA fragments, produced after post-transcriptional cleavage, *cpn10* mRNA is translated more actively than that of *cpn60.1* mRNA (Figure 4.1 &Figure 4.3) favouring the hypothesis that *cpn60.1* mRNA might degrade rapidly after post-transcriptional cleavage event. Besides, the transcription and translation levels of *cpn60.2* are much higher than *cpn60.1* but similar to the *cpn10*; again, this supports the idea that despite being at separate locations on the chromosomes, the levels of expression of *cpn10* and *cpn60.2* genes match each other (Figure 4.2 & Figure 4.4).

In summary, studies on both *M. tuberculosis* and *M. smegmatis* show that the expression and relative levels of induction of *cpn* genes are not the same. Generally, *cpn10* is highly expressed, followed by *cpn60.2*. *cpn60.1* is the least expressed gene both under normal and stressed conditions, indicating additional levels of regulation which might include mRNA processing events leading to differential translation or degradation of the *cpn10-cpn60.1* mRNA.

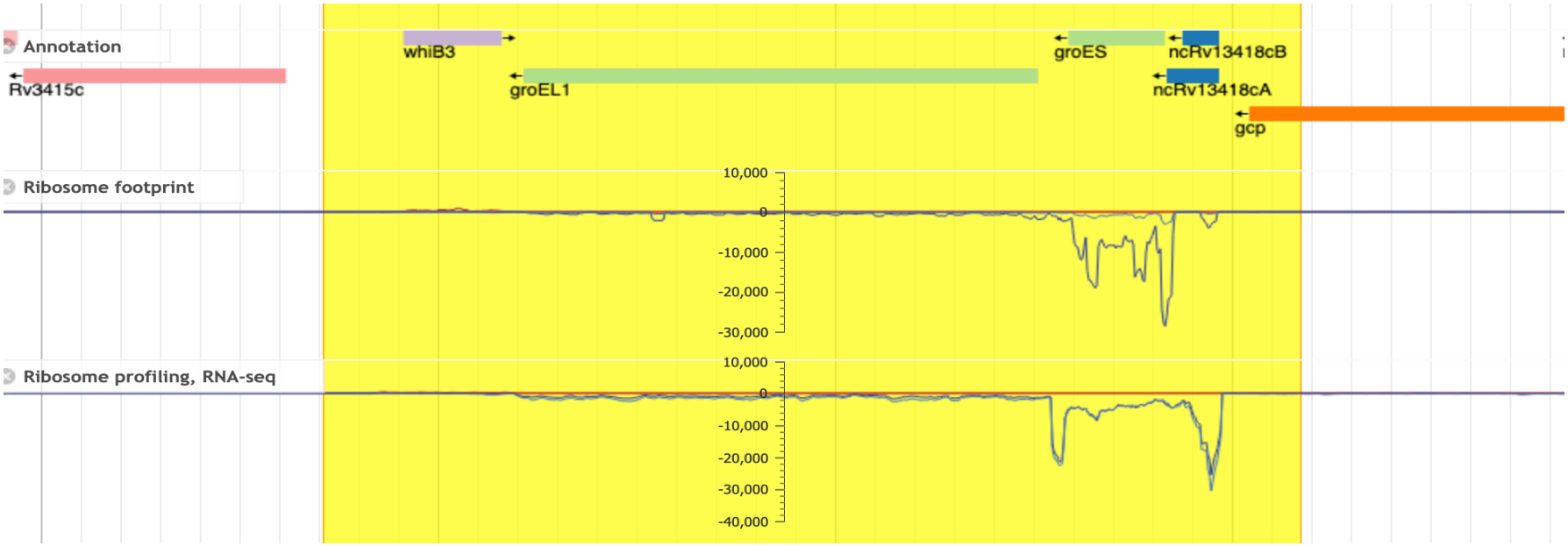


Figure 4.1 Ribosome profiling and RNA-seq graphs of *cpn10* and *cpn60.1* of *M. tuberculosis*.
A screen capture of Mycobrowser portal (https://mycobrowser.epfl.ch/) showing the ribosome profiling and RNA-seq data of *cpn10* (labelled as *groES*) and *cpn60.1* (labelled as *groEL1*) genes of *M. tuberculosis*.

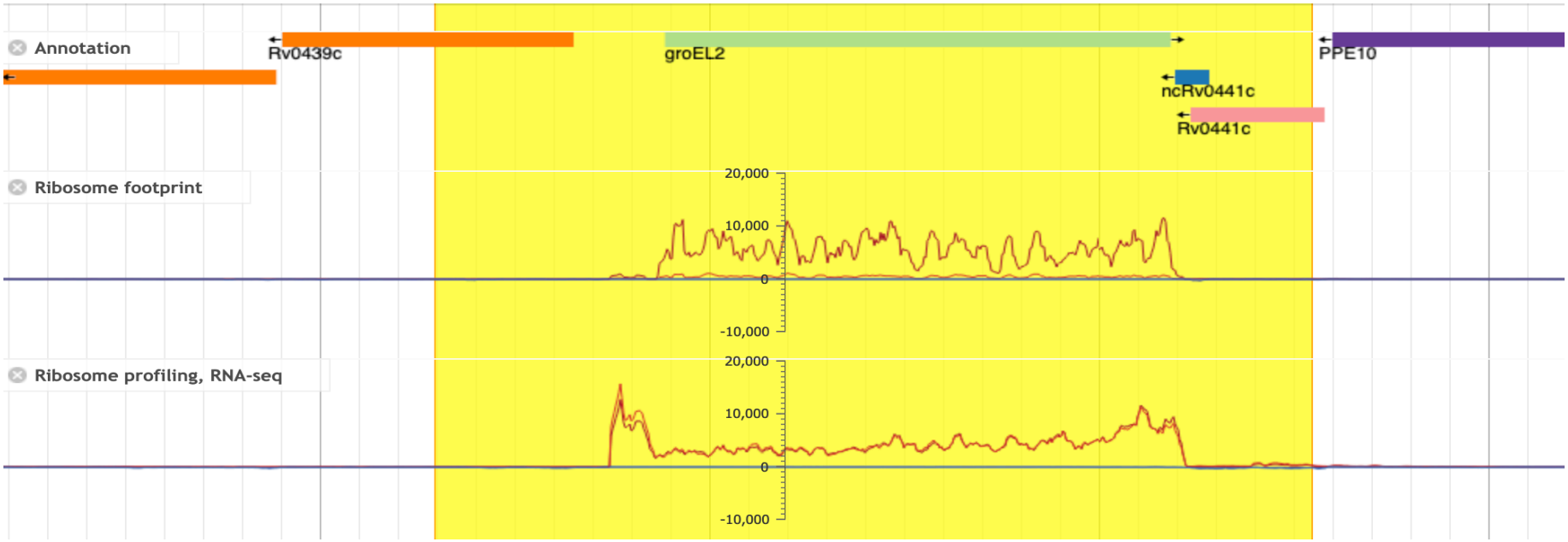


Figure 4.2 Ribosome profiling and RNA-seq graphs of *cpn60.2* of *M. tuberculosis*.
A screen capture of Mycobrowser portal (https://mycobrowser.epfl.ch/) showing the ribosome profiling and RNA-seq data of *cpn60.2* (labelled as *groEL2*) gene of *M. tuberculosis*.



Figure 4.3 Ribosome profiling and RNA-seq graphs of *cpn10* and *cpn60.1* of *M. smegmatis*.
A screen capture of Mycobrowser portal (https://mycobrowser.epfl.ch/) showing the ribosome profiling and RNA-seq data of *cpn10* (labelled as *groS*) and *cpn60.1* (labelled as *groL2*) gene of *M. smegmatis*.

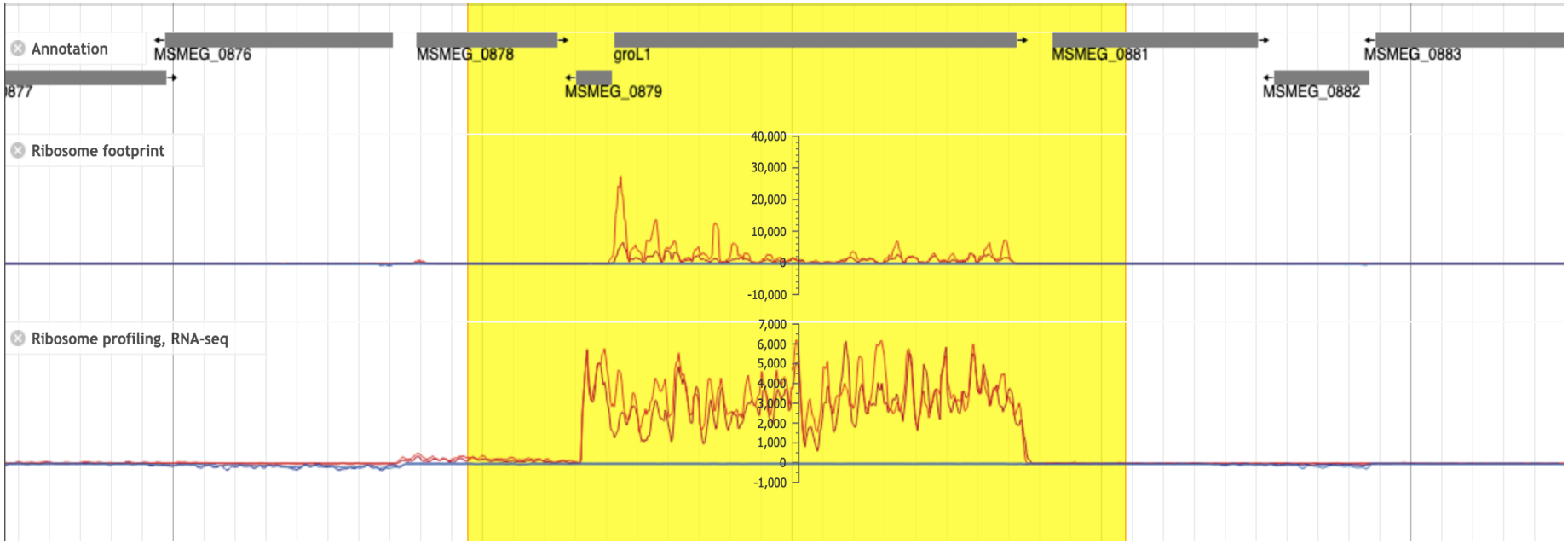


Figure 4.4 Ribosome profiling and RNA-seq graphs of cpn60.2 of *M. smegmatis*.
A screen capture of Mycobrowser portal (https://mycobrowser.epfl.ch/) showing the ribosome profiling and RNA-seq data of *cpn60.2* (labelled as *groL1*) gene of *M. smegmatis*.

As described above, studies have shown that despite being separate in the chromosome, *cpn10* and *cpn60.2* share similar expression levels. This might be due to the reason that these two genes are regulated by a similar mechanism and that the two proteins work together. The following sections will provide some details about the studies done to understand the regulatory mechanism of heat shock genes in bacteria, especially the chaperonin genes in mycobacteria.

4.1.4 Regulation mechanism of heat-shock genes in bacteria

In general, induction and repression are the two heat shock regulation mechanisms described in bacteria (Zuber & Schumann, 1994). *E. coli* is often used as a paradigm in eubacteria in general. Its heat shock expression regulation is well understood, but studies on other bacteria show that it not widely utilised. In *E. coli*, Upon heat shock from 30°C to 42°C, the intracellular concentration of an alternate sigma factor σ^{32} (encoded by *rpoH* gene) rises 15 to 20-fold within 5 minutes of the heat shock and then drops to become steady to a new level. This happens due to two reasons; one is the increased translation of *rpoH* mRNA and second is the enhanced stability of σ^{32} due to change in the rate of σ^{32} degradation (mostly by the FtsH protease). σ^{32} directs RNA polymerase to the promoters of heat-shock genes (e.g., *groESL* and *dnaK*), thereby increasing the transcription levels of chaperonin genes. On the other hand, σ^{32} must be bound to DnaK to be acted on by FtsH. When there is an increase in unfolded proteins, like after heat shock, DnaK binds to these, releasing σ^{32} so it is no longer degraded. Thus, this system responds both to temperature (via altered translation) and to unfolded proteins. (Grossman et al., 1987; Roncarati & Scarlato, 2017; Wang & DeHaseth, 2003).

Studies on *B. subtilis* and subsequently many other bacteria showed that the use of alternate sigma factors to regulate heat shock genes is not universal. In *B. subtilis* a 9-bps inverted repeat (IR) sequence separated by 9 bps (TTAGCACTCTTTAGTGCTGAGTGCTAA) was identified in the regulatory regions of *dnaK* and *groESL* genes. This sequence is named "Controlling Inverted Repeat of Chaperone Expression" (CIRCE). It is interesting to note that the distance between the IRs is usually equal to 9-bps which is equivalent to one turn of a helix (Zuber & Schumann, 1994). Because the deletion or mutation of CIRCE in both the *groE* and *dnaK* promoters of *B. subtilis* resulted in overexpression of both the operons, it was suggested that CIRCE functions as an operator (Yuan & Wong, 1995).

Wong, 1995 demonstrated that in *B. subtilis*, *orf39* (the first gene of *dnaK* operon in *B. subtilis*) encodes a negative regulatory protein which acts as a CIRCE-specific repressor for *dnaK* and *groE* operons. Further studies showed that in *B. subtilis*, the deletion of *orf39* gene (now named as *hrcA*; encoding a 39kDa protein HrcA) caused the constitutive upregulation of both the *groESL* and *dnaK* operons at normal growth

temperatures. HrcA stands for "Heat Regulation at CIRCE". Hence, in *B. subtilis*, HrcA repressor is known to co-regulate both the *dnaK* and *groESL* family heat shock genes (Hecker et al., 1996). Negative regulation mechanism of heat shock promoters was first observed in this organism (Narberhaus, 1999).

Some studies also demonstrated a relationship between HrcA and GroEL protein in *B. subtilis*. Purified HrcA protein bound to DNA containing CIRCE sequence; the effect increased considerably by the addition of GroEL. It was therefore proposed that GroEL binding with HrcA is necessary to bring the repressor into its active, inhibitory conformation. Upon heat shock chaperonins are engaged in the necessary chaperone activities, hence are not available for HrcA binding rendering it inactive (Minder et al., 2000; Mogk, 1997; Narberhaus & Bahl, 1992). However, we don't know the exact mechanism of dissociation of HrcA upon heat shock and the interaction of HrcA with the CIRCE element and GroEL is still being investigated. Based on the above knowledge, (Figure 4.5 & Figure 4.6) show how HrcA is proposed to regulate chaperonin genes under normal and stressed conditions. In both the heat shock mechanisms, σ^{32} and HrcA respond to unfolded proteins, even though they use different ways to do this.

In Streptomyces, the CIRCE sequence is found only in the upstream regions of chaperonin genes and not in that of *dnaK*, whereas another IR sequence CTTGAGTCAGCTTCACTCAAG was found in the upstream region of *dnaK* (Grandvalet et al., 1999), indicating two different controlling mechanisms for this organism. Studies have identified another repressor called HspR. Under normal growth conditions, HspR (encoded by *hspR*) binds to the IR sequence present in the promoter region of the *dnaK* operon in *Streptomyces albus* causing its repression. Upon heat shock, this binding is removed causing overexpression of *dnaK* operon (Grandvalet et al., 1997; Servant & Mazodier, 2001). The IR sequence in the *dnaK* regulatory locus, regulated by HspR, is termed as HAIR (HspR-associated inverted repeat) and this repression mechanism is termed as HAIR/HspR repression system (Jäger et al., 2004). On the other hand, HrcA was found to negatively regulate only the chaperonin genes (*cpn10-cpn60.1* and *cpn60.2*) in *Streptomyces albus* and not the *dnaK* operon (Grandvalet et al., 1998) . CIRCE sequences comparable to the CIRCE sequence of *B. subtilis* have been reported in *Streptomyces* and *Mycobacteria* (Servant & Mazodier, 2001; Wetzstein et al., 1992). Based on the evidence above, it was hypothesized that HrcA mainly regulates the CIRCE

containing operons while HspR mainly binds to the HAIR sequences (Stewart et al., 2002).

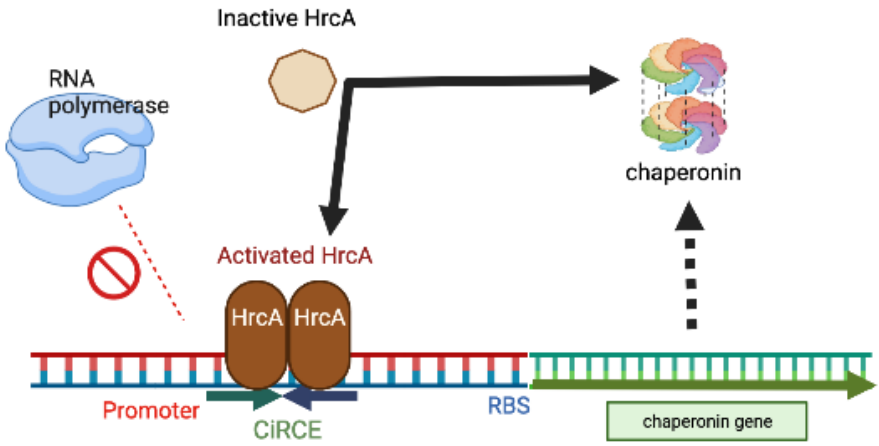


Figure 4.5 HrcA binding with the CIRCE element under normal conditions.
Under normal growth conditions, HrcA binds with the CIRCE element causing repression of the chaperonin operon. The repression is not static, and a steady level of chaperonin is always maintained in the cell. The model is adapted from (Mogk, 1997; Narberhaus, 1999).

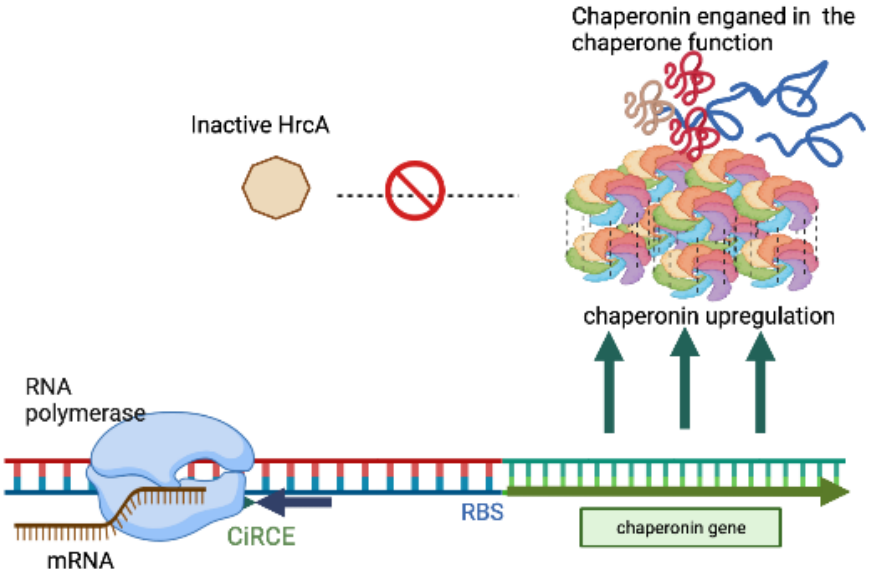


Figure 4.6 HrcA binding with the CIRCE element following heat-shock.
Following heat shock, HrcA dissociates from the CIRCE sequence leading to the unrestricted binding of RNA polymerase with the promoter resulting in overexpression of the chaperonin genes. The model is adapted from (Mogk, 1997; Narberhaus, 1999).

4.1.5 Regulation mechanism of heat-shock genes in *Mycobacteria*

A targeted mutagenesis study demonstrated that the knockout of *hspR* gene (Δ*hspR)* showed no significant effect on the expression of chaperonin genes in *M. tuberculosis* (Stewart et al., 2001) except a small increase of 1.96-fold in the expression of *cpn10* compared to the W.T (Stewart et al., 2002). However, the double mutant (Δ*hspR*Δ*hrcA)* showed 9.60, 4.60 and 2.41-fold upregulation in the expression of *cpn10*, *cpn60.2* and *cpn60.1* genes respectively: this indicates that the predominant regulator for the chaperonin genes in *M. tuberculosis* is HrcA (Stewart et al., 2002). Two putative HAIR elements were found upstream of *dnaK* gene in *M. tuberculosis* GTTGAGTGCATCAGGCTCAGC and CTTGAGCGGGGTGCACTCATC (Parijat & Batra, 2015), whereas two CIRCE sequences GCACTCGCGACCGGTGAGTGC and GCACTCGGCATAGGCGAGTGC were found upstream of *cpn60.2* and one CIRCE sequence GCACTCTCATGTATAGAGTGC was found upstream of *cpn10* gene. This again is consistent with the hypothesis that HrcA is the major regulator of chaperonin family genes (Stewart et al., 2002).

HrcA binds to the CIRCE element and reduces the level of transcription in unstressed condition. However, the CIRCE sequences are not constantly occupied by HrcA, and this repression is not 100% static; hence, it does not completely block the expression of the *cpn* genes, a leaky expression of these genes is always noticed (Narberhaus & Bahl, 1992; Schön & Schumann, 1993) . Besides, it is not known how HrcA binds to the CIRCE element, it was hypothesized that perhaps HrcA binds separately to both sides (left and right) of a CIRCE sequence (Zuber & Schumann, 1994).

It could also be worth to notice that two CIRCE elements are found in the *cpn60.2* regulatory region while only one CIRCE element is present in the upstream region of *cpn10*; this might cause tighter repression of *cpn60.2* compared to *cpn10* which in turn could be a possible reason for why *cpn10* is expressed at higher levels than *cpn60.2*.

Moreover, heat shock response is not only a convenient method for understanding the expression regulation but also a significant element in the pathogenesis of *Mycobacterium tuberculosis*. Following infection, the heat shock proteins are upregulated in *M. tuberculosis*; this might help in the survival of the bacilli in the stressful phagocytic environment. Fractional interruption of heat shock regulation in

M. tuberculosis impairs the ability of the bacillus to cause persistent infection (Stewart et al., 2001). HrcA/CIRCE heat shock regulation mechanism is not fully understood in mycobacteria. A further investigation and understanding might prove to be helpful in finding novel approaches to combat chronic tuberculosis infection (Stewart et al., 2002).

No previous work has been reported about the promoter activity and expression analysis of the chaperonin genes in *M. marinum*. The next sections of this chapter will cover the bioinformatics analysis of the upstream regions of the chaperonin genes in actinobacteria and a few other eubacteria. In the later sections the experimental plan to determine the promoter regions of the chaperonin genes of *M. marinum* will be discussed. This was done by the construction of promoter probe vectors. And the results of the qualitative and quantitative assays to measure the promoter activity using the recombinant plasmids will also be discussed in the last section.

4.2 Bioinformatics Analysis

Promoters are essential DNA elements located in the regulatory region of an operon, upstream of the coding region, and are responsible for the regulation of gene expression by controlling the transcription initiation (Silva et al., 2011). These are short sequences, approximately 60 to 120 bps, usually located around the transcription start site (TSS) (Oubounyt et al., 2019). In prokaryotes, the promoter region usually contains two consensus sequences: the Pribnow – a sequence of TATAAT located nearly at -10 position of the TSS, a -35 element – a sequence of TTGACA (15 to 21 bps upstream of the Pribnow sequence) (Hawley & McClure, 1983; He et al., 2018). Promoter regions of a genome also have a different stability and curvature pattern than the coding regions; they are usually less stable and more bendable. Promoters control the transcription initiation primarily by recognition and interaction with the σ factor subunit of RNA polymerase. Different σ factors recognize and bind with different promoters depending on consensus sequence motifs and their characteristics (Lai et al., 2019). Apart from the regulation of gene expression, these regulatory regions also play their roles in stability of the mRNA transcripts, and translation. In short, the coding sequences of an operon are analogous to a computer memory whereas the promoter regions represent a computer program that acts on the memory (Howard & Benson, 2003; Kanhere & Bansal, 2005).

Mycobacteria are G+C rich (62-70%) acid fast bacilli with exceptionally low transcription rate (Cook et al., 2009). Since G+C content of a genome affect the gene expression and regulation by codon usage bias and recognition of promoter regions, it is therefore observed that in most cases, the mycobacterial promoters do not function efficiently in *E. coli* (Agarwal & Tyagi, 2006). Studies have shown that majority of mycobacterial promoters possess sequences resembling the Pribnow box in *E. coli* but only a few have a comparable -35 consensus sequence (Kalate et al., 2003). It shows the presence of some additional sequences in the promoter regions required for the regulation of gene expression in mycobacteria. As discussed earlier in sections 4.1.4 and 4.1.5, the inverted repeat sequences (CIRCE and HAIR) play an important role in the regulation of expression of heat-shock genes in *Bacillus* and mycobacterial species by binding to the HrcA and HspR repressors.

Moreover, as described earlier in section 4.1.3, despite being in a single operon the lower expression of *cpn60.1* compared to *cpn10* indicates some additional level of regulation. This could be because bacteria have very remarkable energy efficient machinery that rapidly reshapes the proteome by different regulatory steps: post-transcriptional regulation allows the fastest adaptation than any other type of regulatory mechanism (Chełkowska-Pauszek et al., 2021). mRNA is not only responsible for conveying the genetic information from DNA to ribosomes for protein synthesis but also carry additional layer of information about its own stability and the rate and extent of translation such as the frequency of translation initiation, the speed of elongation and the folding of the newly formed polypeptide (Lorenz et al., 2016). Particularly, the intrinsic ability of RNA to fold into dynamic secondary and tertiary structures making stem-loop, hairpins, and pseudoknots can have a strong impact on post-transcriptional events (Chiaruttini & Guillier, 2019; Wen et al., 2008). It is therefore hypothesized that the formation of mRNA secondary structures in the intergenic region of *cpn10-cpn60.1* might cause post-transcriptional cleavage and subsequent degradation of *cpn60.1* mRNA or it leads to translation disruption resulting in the lower yield of Cpn60.1.

Furthermore, we don't have enough information about HrcA/CIRCE interaction. Some studies demonstrated that probably HrcA dimer binds with both arms of the CIRCE sequences. In literature, the only crystal structure study of HrcA repressor was performed using *Thermotoga maritima* HrcA, which crystalized as a dimer and was found to be incompatible with DNA binding which might represents an inactive form of the repressor (Liu et al., 2005). Mutagenesis studies showed the three conserved amino acids at the N-terminal of the HrcA might be involved in binding with DNA. Two of these were studied in *B. subtilis*, where a double mutation in the two amino acids S39G and A40G resulted in the complete loss of binding of HrcA with DNA (Wiegert & Schumann, 2003). While in another study the mutation in the amino acid R43D in *Bacillus thermoglucosidasius* resulted in considerable loss of HrcA binding with DNA (Hitomi et al., 2003). However, surprisingly the sequence identity of HrcA repressor from different bacterial genera is usually less than 30% (Narberhaus, 1999) indicating that the molecular characters of different HrcA repressors might not corelate with each other.

In the following sections, I will present a bioinformatics analysis of the upstream regions of *cpn* genes in *M. tuberculosis* and *M. marinum* compared to 50 mycobacterial species and 12 other actinobacteria. The purpose is to investigate the conserved consensus sequences in the upstream regions of *cpn* genes and predict the secondary structures in the *cpn10-cpn60.1* intergenic region. Also, the alignment and identity of HrcA repressor of mycobacteria with that of *Bacillus* will be presented and discussed.

4.2.1 Bioinformatics analysis of promoter regions of *cpn* genes

It was observed that, in *M. marinum*, the upstream region of *cpn10* has one CIRCE element with a 9-bps spacer located three nucleotides upstream of the predicted TSS (GTG). A potential -10 element (Pribnow box) was found in the spacer region of CIRCE and -35 sequence at 15 bps upstream of the -10. An inverted repeat sequence with a spacer of 29 bps was also detected in the intergenic region of *cpn10-cpn60.1* (Figure 4.7). Whereas, in the upstream region of *cpn60.2* two CIRCE sequences were observed (Figure 4.8). Most of the upstream sequence is G+C rich (>50%) except the small region where Pribnow box is present (Figure 4.9 &Figure 4.10).

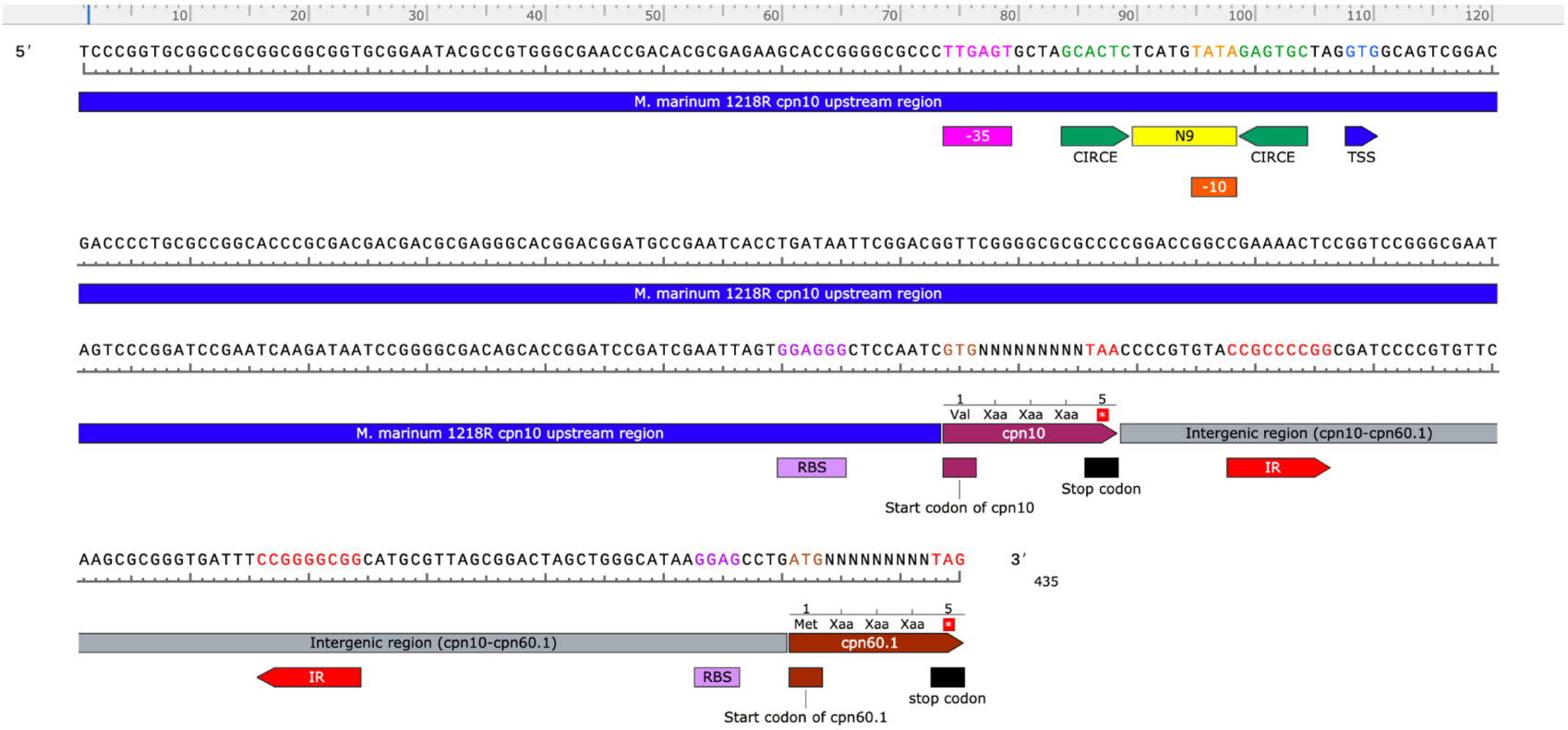


Figure 4.7 Upstream region of *M. marinum cpn10* and intergenic region of *cpn10-cpn60.1*.
CIRCE sequence, potential TSS, TATA box (-10) and -35 elements are highlighted. The IR sequence in the intergenic region is marked. The coding regions for *cpn10* and *cpn60.1* are not shown, but just designated as NNNNNNNNN.

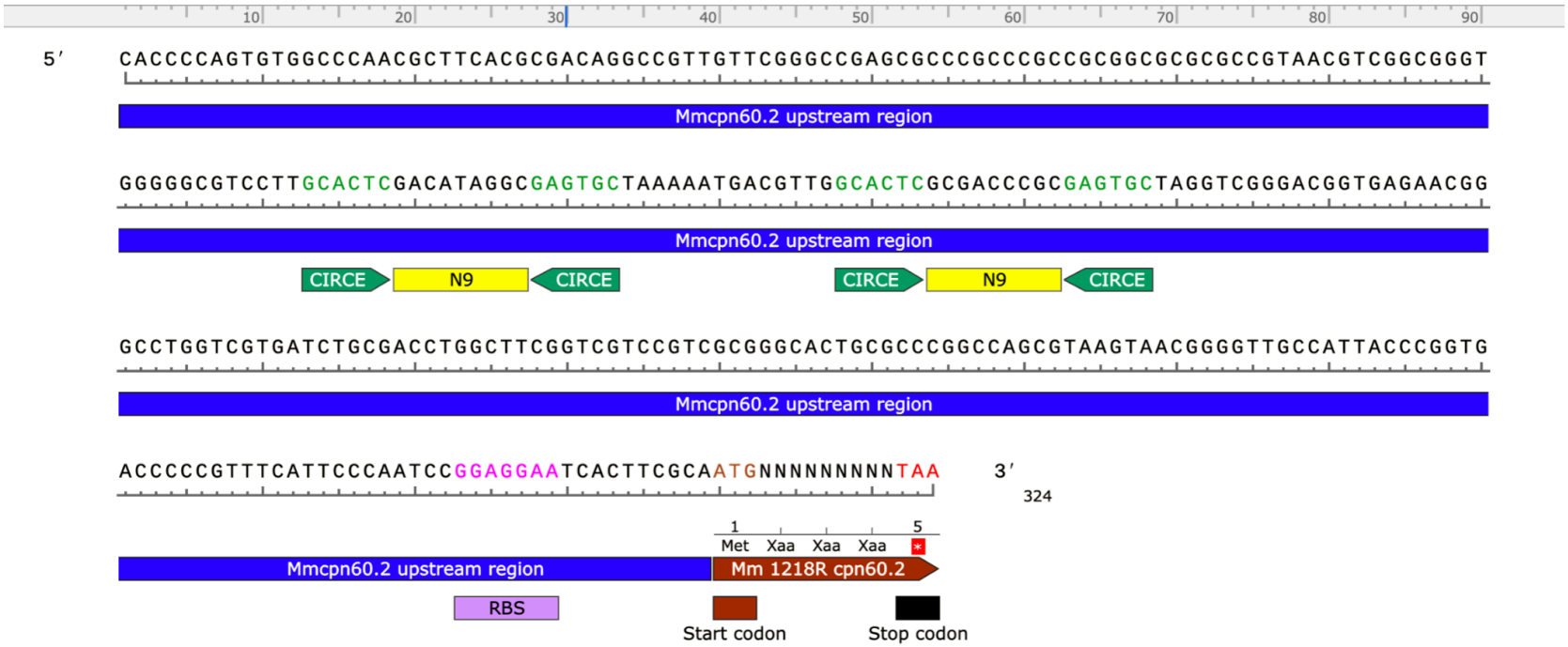


Figure 4.8 Upstream region of *M. marinum cpn60.2*.
CIRCE sequences are highlighted. The coding region for *cpn60.2* is not shown, but just designated as NNNNNNNNN.

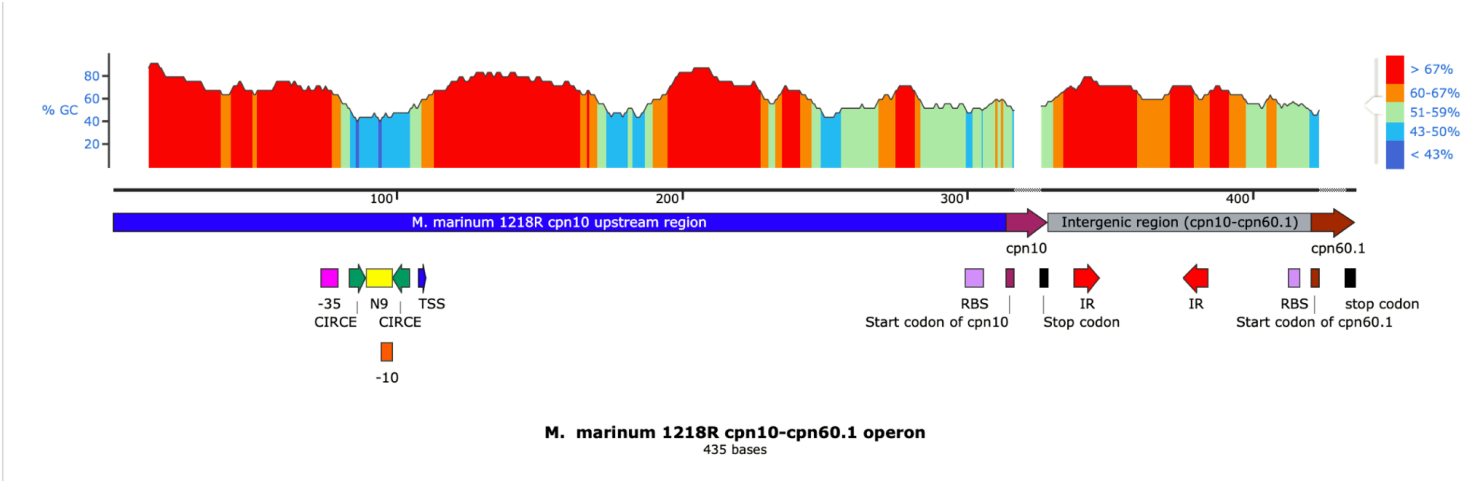


Figure 4.9 Upstream region of *M. marinum cpn10* and intergenic region of *cpn10-cpn60.1* showing G+C percentage.

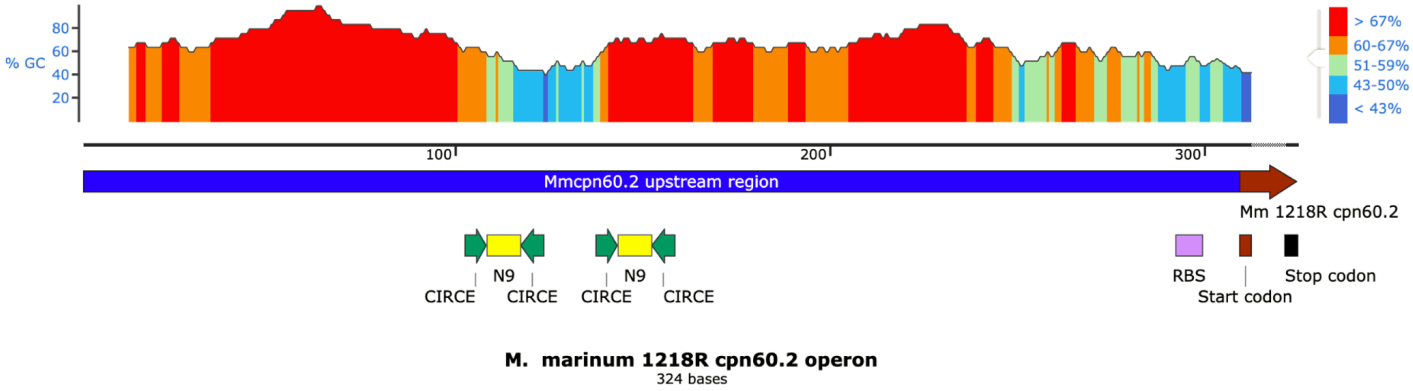


Figure 4.10 Upstream region of *M. marinum cpn60.2* showing G+C percentage.

Whole genome sequences of 50 species of mycobacteria including *M. marinum, M. smegmatis, M. tuberculosis and M. bovis* and 12 species of other actinobacterial genera were downloaded from NCBI. The upstream regions of *cpn* genes were copied and used for the alignment using default parameters of MUSCLE (Multiple **S**equence Comparison by Log- Expectation). The sequence alignments showing the conserved regions and CIRCE elements are displayed in Figure 4.11–Figure 4.16. In the upstream region of *cpn10,* both arms of the CIRCE element GCACTCN9GAGTGC were found to be highly conserved (without a single mismatch in the arms region) among all the 50 species of mycobacteria. Out of 50 species, in 25 species such as *M. marinum, M. smegmatis, M. tuberculosis and M. bovis* not only the CIRCE arms but the 9-bps spacer region TCATGTATA was also found to be conserved. In the rest of the 25 species the first and third bases T and A are replaced by either G or C (Figure 4.11). However, the significance of the conservation of the spacer region is still not understood. When 12 other species of actinobacteria belonging to the genera *Streptomyces, Corynebacterium, Micrococcus*, *Amycolatopsis, Nocardioides, Cutibacterium, Rhodococcus* and *Micromonospora* were included in the analysis, the CIRCE sequence was observed to be conserved among all species except a couple of mismatches. In the right arm of the CIRCE sequence, A was observed to be replaced with G in *M. aurantiaca*, *N. rotundus* and *M. luteus* while in *C. acnes* A was replaced with T. No mismatch was seen in the left arm of the CIRCE sequence, except in *C. pseudotuberculosis* where the last C of the left arm is replaced by T. Considerable diversity was seen in the 9-bps spacer region among all the 12 actinobacterial genera used in the analysis (Figure 4.12). However, I could not find any published studies about the importance of the spacer region in these promoters. There are two possibilities for this (a) the spacer sequence may be particularly important for some reason in mycobacteria, hence its high conservation. This would imply either that it's bound by something (unlikely) or that it gives very precise physical properties to the region. (b) It might simply be a reflection of the relatively recent separation of the different Mycobacterial species, there may not have been time for the sequences to diverge.

Three bps upstream of the right arm of the CIRCE, the TSS was marked, and the Prinbnow box was observed to be 100% conserved at the -10 site of the predicted TSS in all the 50 species of mycobacteria included in this analysis. The -35 element was also

conserved in all the mycobacterial species used. However, in the 12 other species of actinobacteria remarkable diversity was seen in these elements (Figure 4.11 & Figure 4.12). Nearly 28 bps downstream of the TSS, another highly conserved sequence (GC box) 5' CCCGCGACG 3' was found in all the 50 species of mycobacteria (Figure 4.13).

The alignment of the upstream region of *cpn60.2* showed 100% conservation of both arms of the two CIRCE sequences in all the 50 species of mycobacteria. Both the CIRCE sequences were found to be conserved in 12 other species of actinobacteria except a couple of point mismatches in right arms. Similar to the GC box in the upstream of *cpn10*, a highly conserved GC box 5' CGCGGGC 3' was found downstream of the CIRCE sequences in all the 50 species of mycobacteria. No conserved -10 or -35 element was observed from the alignment. These elements are still to be investigated in the promoter regions of the *cpn* genes (Figure 4.14 –Figure 4.16). I could not find any information in the published studies about the significance of the conserved GC boxes present in the regulatory regions of the mycobacterial operons. The high level of conservation shows the importance of these regions in the selection pressure during evolution.

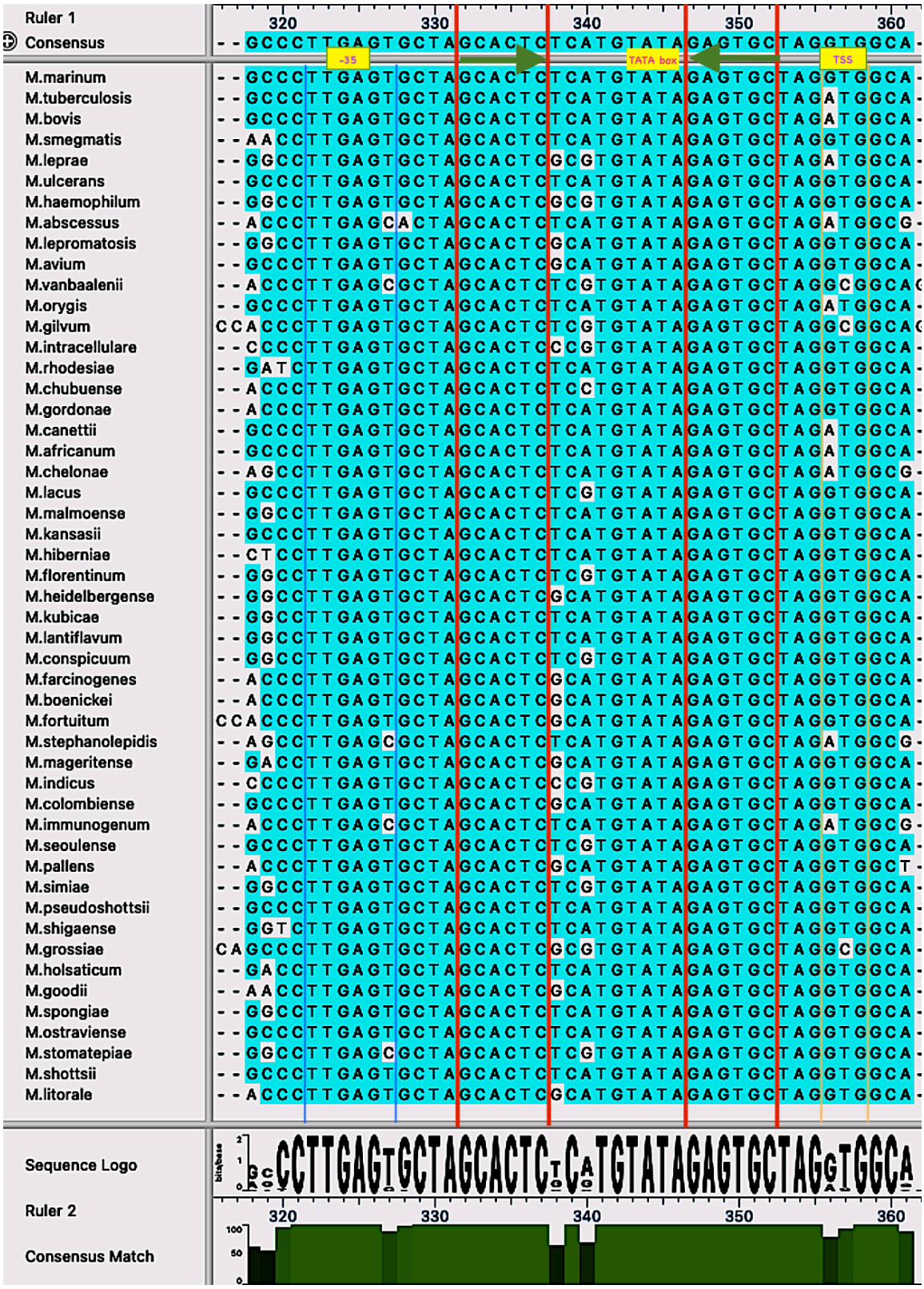


Figure 4.11 Alignment of upstream region of *cpn10* gene of 50 species of mycobacteria. The highly conserved CIRCE sequence is marked in red lines. Potential TSS, TATA box and -35 element are also marked.

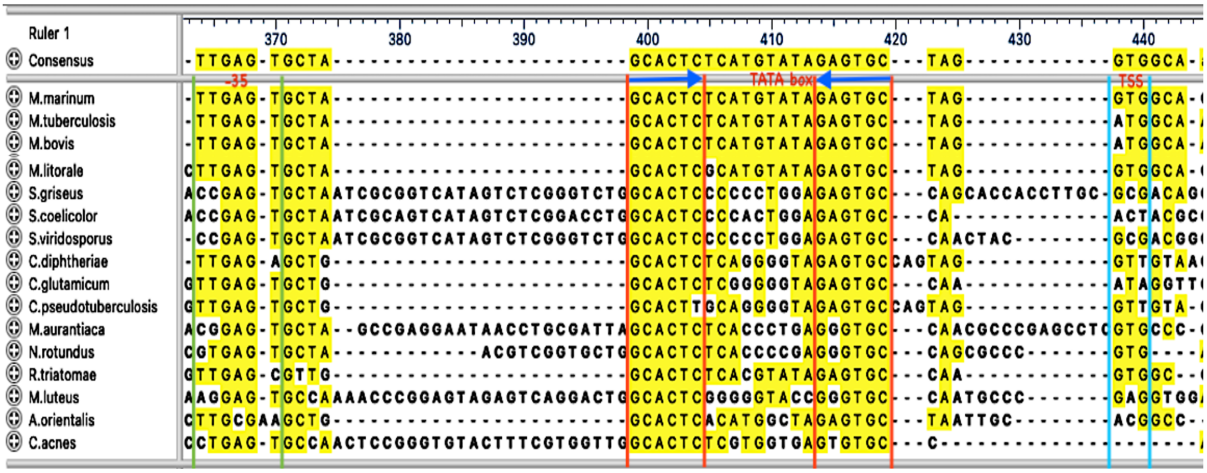


Figure 4.12 Alignment of upstream region of *cpn10* gene of 12 species.
Alignment of upstream region of *cpn10* gene of 12 species of actinobacteria with some representatives of mycobacteria. The consensus CIRCE sequence is marked in red lines. Potential TSS, TATA box and -35 element are also marked.

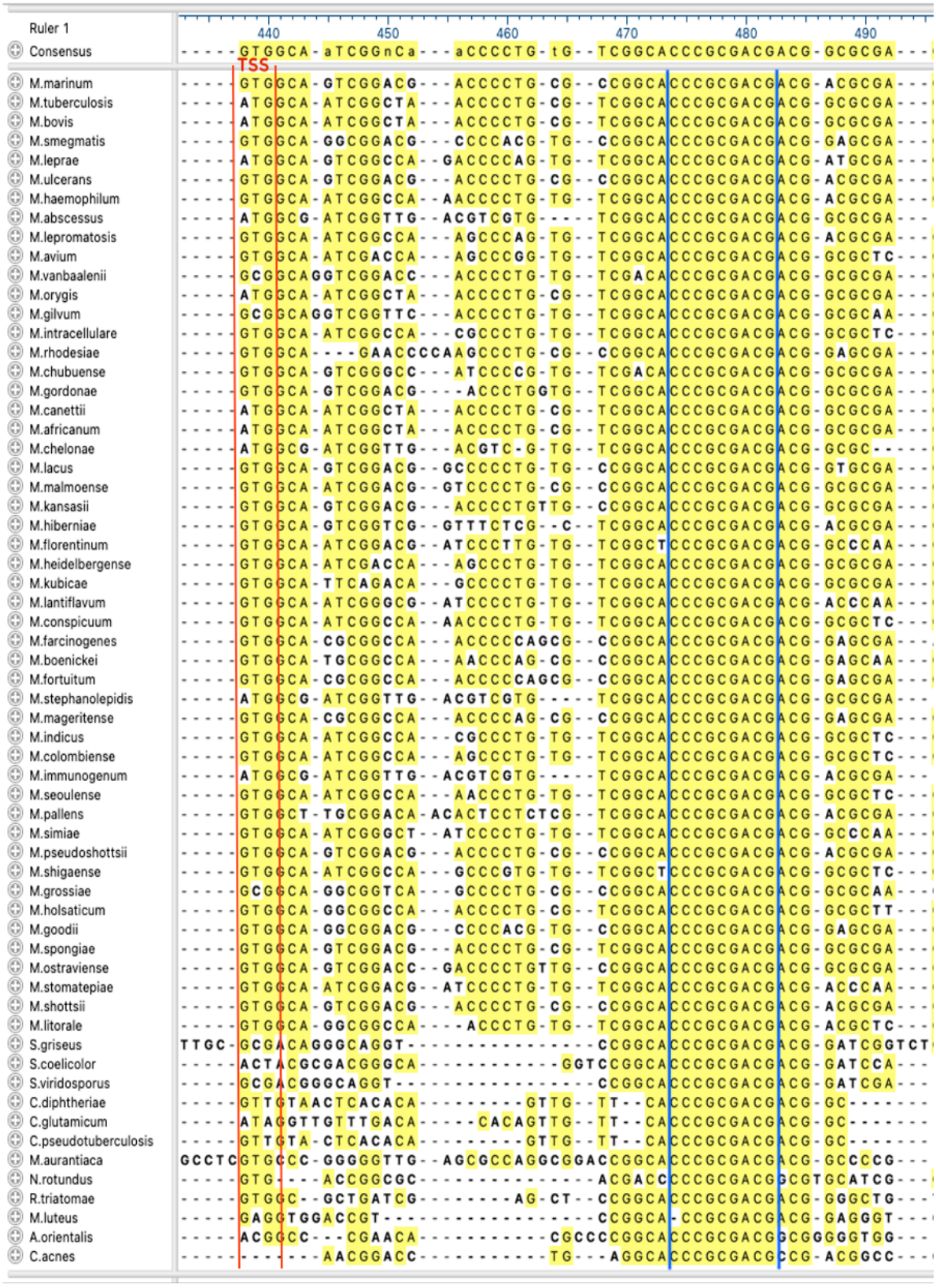


Figure 4.13 Alignment of upstream region of *cpn10* gene of 62 species.
Alignment of upstream region of *cpn10* gene of 50 species of mycobacteria and 12 other species of actinobacteria. The highly conserved GC box downstream of the TSS is marked in blue lines.

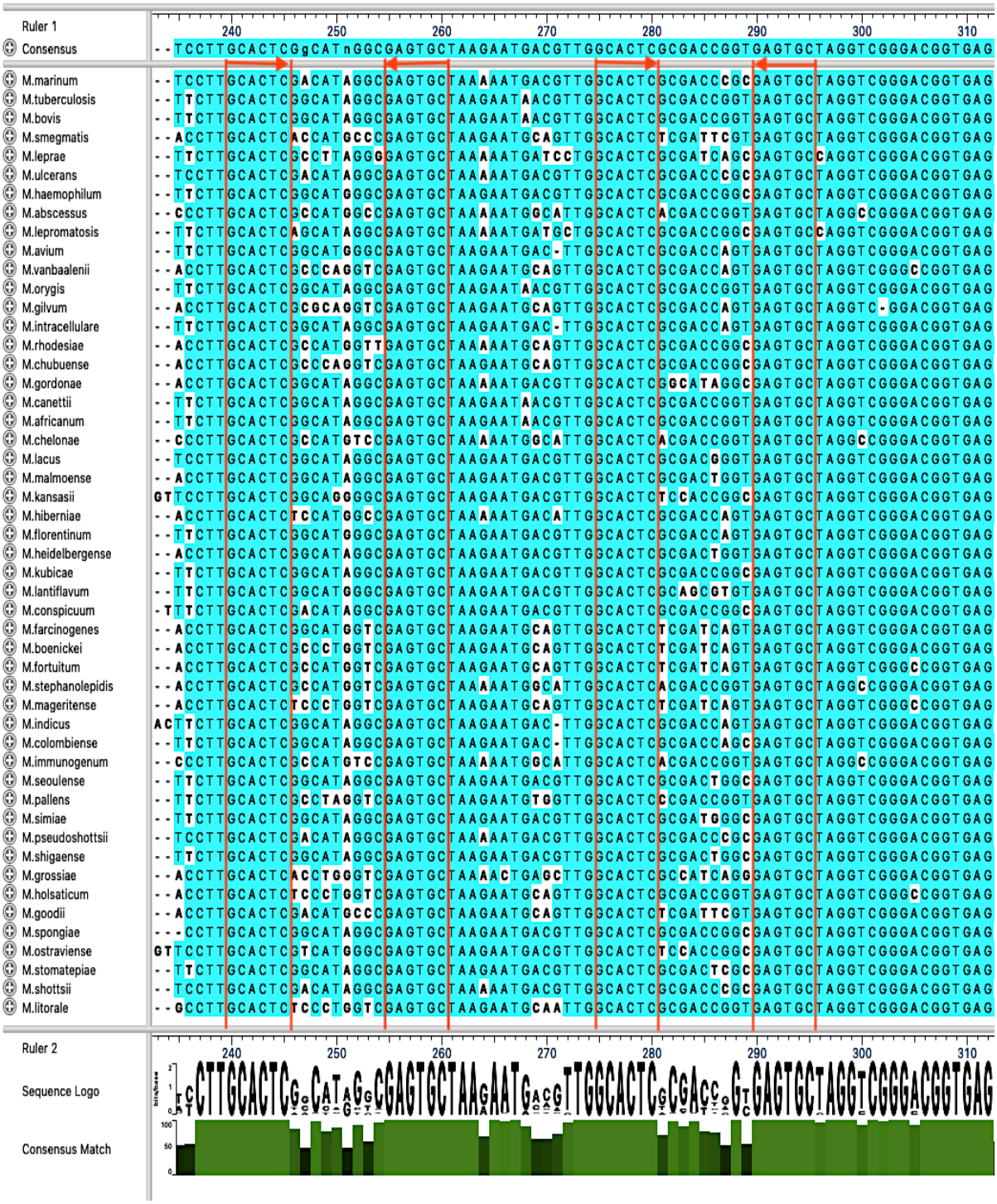


Figure 4.14 Alignment of upstream region of *cpn60.2* gene of 50 species of mycobacteria. The two highly conserved CIRCE sequences are marked in red lines.



Figure 4.15 Alignment of upstream region of *cpn60.*2 gene of 12 species.
Alignment of upstream region of *cpn60.2* gene of 12 species actinobacteria with some representatives of mycobacteria. The two consensus CIRCE sequences are marked in red lines.

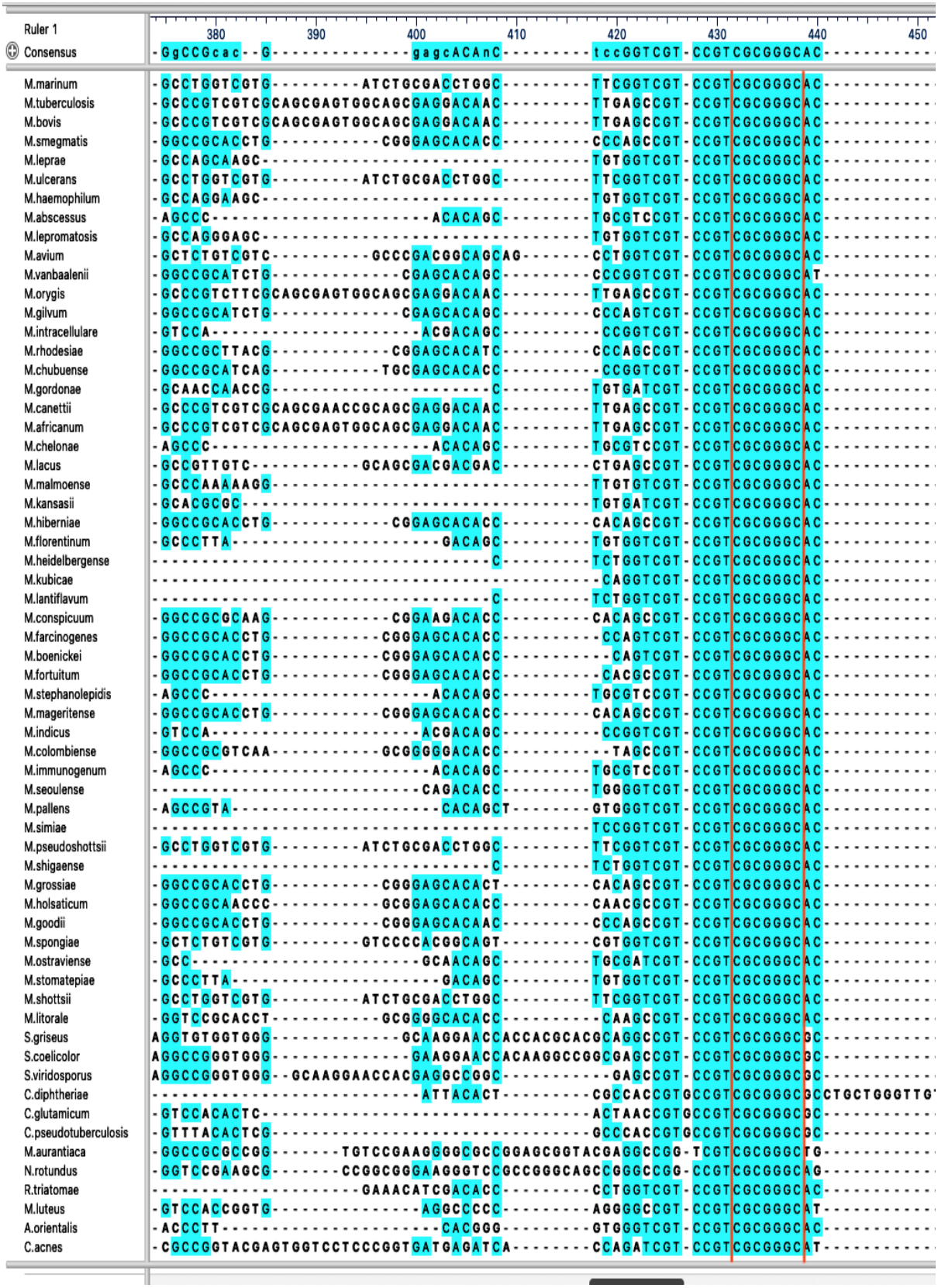


Figure 4.16 Alignment of upstream region of *cpn60.2* gene of 62 species.
Alignment of upstream region of *cpn60.2* gene of 50 species of mycobacteria and 12 other species of actinobacteria. A highly conserved GC box downstream of the CIRCE sequences is marked in red lines.

4.2.2 Analysis of mRNA secondary structures in the *cpn10-cpn60.1* intergenic region

The *groES-groEL* intergenic region in *E. coli* MG1655 is only 43 bps while intergenic region of *cpn10-cpn60.1* in *M. tuberculosis* H37Rv and *M. marinum* 1218R is 94 and 92 bps respectively. This long intergenic region within the operon is interesting in the light of the differential regulation of Cnp10 and Cpn60.1 discussed above in section 4.1. The secondary structures formed in the mRNA could alter its translatability or act as a processing site for endogenous RNAase activity. To look for evidence of such structures, the DNA sequences of these three intergenic regions were used to predict the RNA secondary structures using RNAfold (available at https://www.tbi.univie.ac.at/RNA/, (Lorenz et al., 2011) default parameters (Figure 4.17 & Figure 4.18). A strong stem loop structure was found in the predicted secondary structures of both the mycobacterial strains while no such structure was formed in that of *E. coli*. The free energy of the thermodynamic ensemble formed for *E. coli* was only -1.23 kcal/mol while that of *M. tuberculosis* and *M. marinum* was -51.02 kcal/mol and -51.60 kcal/mol respectively.

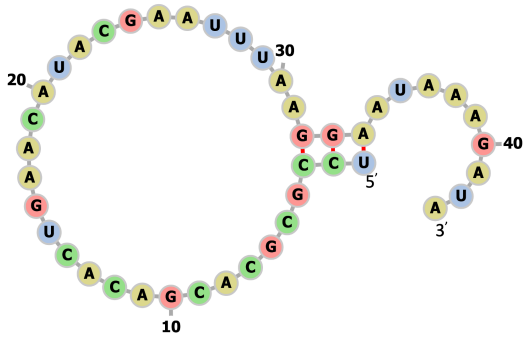


Figure 4.17 MFE structure of *groES-groEL* intergenic region of *E. coli* MG1655.
MFE (Minimum Free Energy) structure (mRNA secondary structure) encoding base-pair probabilities of *groES-groEL* intergenic region of *E. coli* MG1655.

The formation of strong secondary structure in the intergenic region of *cpn10-cpn60.1* of the *M. tuberculosis* and *M. marinum* might indicate the disruption of translation in the intergenic region of bicistronic mRNA or it might have role in post-transcriptional processing. As discussed in the section 4.1.1 and 4.1.2, both in *M. tuberculosis* and *M. smegmatis* the expression level of *cpn60.1* increases by heat shock at 42°C. A possible hypothesis is that the secondary structures formed in the intergenic region cause

mRNA splicing and degradation at normal growth temperature whereas, at heat shock the H-bonding of the mRNA base pairs breaks, leading to the disruption of secondary structure which in turn causes higher expression of *cpn60.1*.

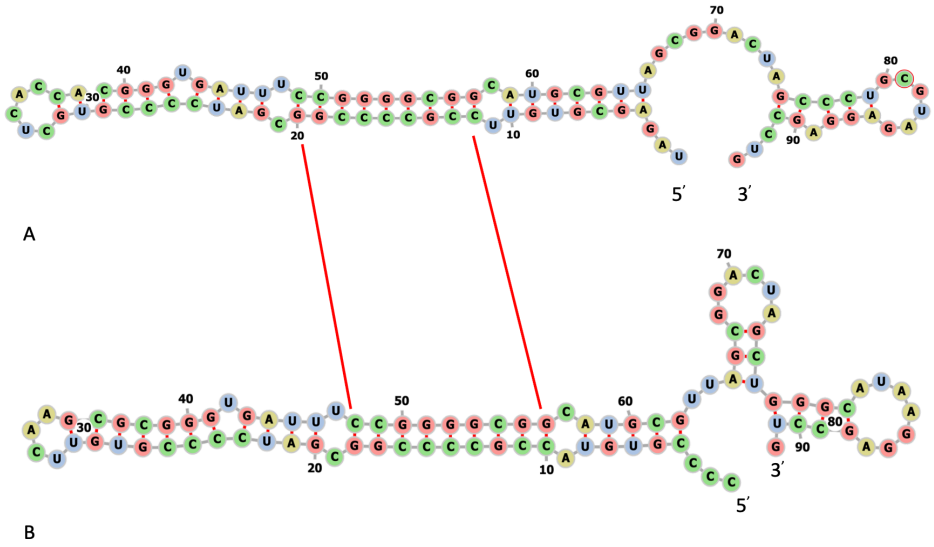


Figure 4.18 The MFE structure encoding base-pair probabilities of *cpn10-cpn60.1* intergenic region.
A) MFE structure (mRNA secondary structure) encoding base-pair probabilities of *cpn10-cpn60.1* intergenic region of *M. tuberculosis* H37Rv. B) Centroid structure drawing encoding base-pair probabilities of *cpn10-cpn60.1* intergenic region of *M. marinum* 1218R. Inverted repeat sequence is shown with red lines.

One conserved GC box inverted repeat was observed in the stem regions of both the MFE (Minimum Free Energy) drawings of *cpn10-cpn60.1* intergenic region of *M. tuberculosis* and *M. marinum*. To investigate the conservation of this IR region in other species of *Mycobacteria*, I performed the sequence alignment of the intergenic regions of *cpn10-cpn60.1* from 47 species of mycobacteria using default parameters of MUSCLE. The sequence alignment is shown in Figure 4.19. The IR sequence was found to be highly conserved among all the 47 species of mycobacteria which indicates the importance of this element in the intergenic region. The alignment sequence was used to predict the mRNA secondary structure using RNAalifold (available at http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAalifold.cgi) default parameters. A similar stem loop structure was found in the predicted secondary structure indicating that the formation of the long stem loop could be universal in all mycobacteria (Figure 4.20). This further shows the importance of the conserved IR in the intergenic region of *cpn10-cpn60.1* in mycobacteria.

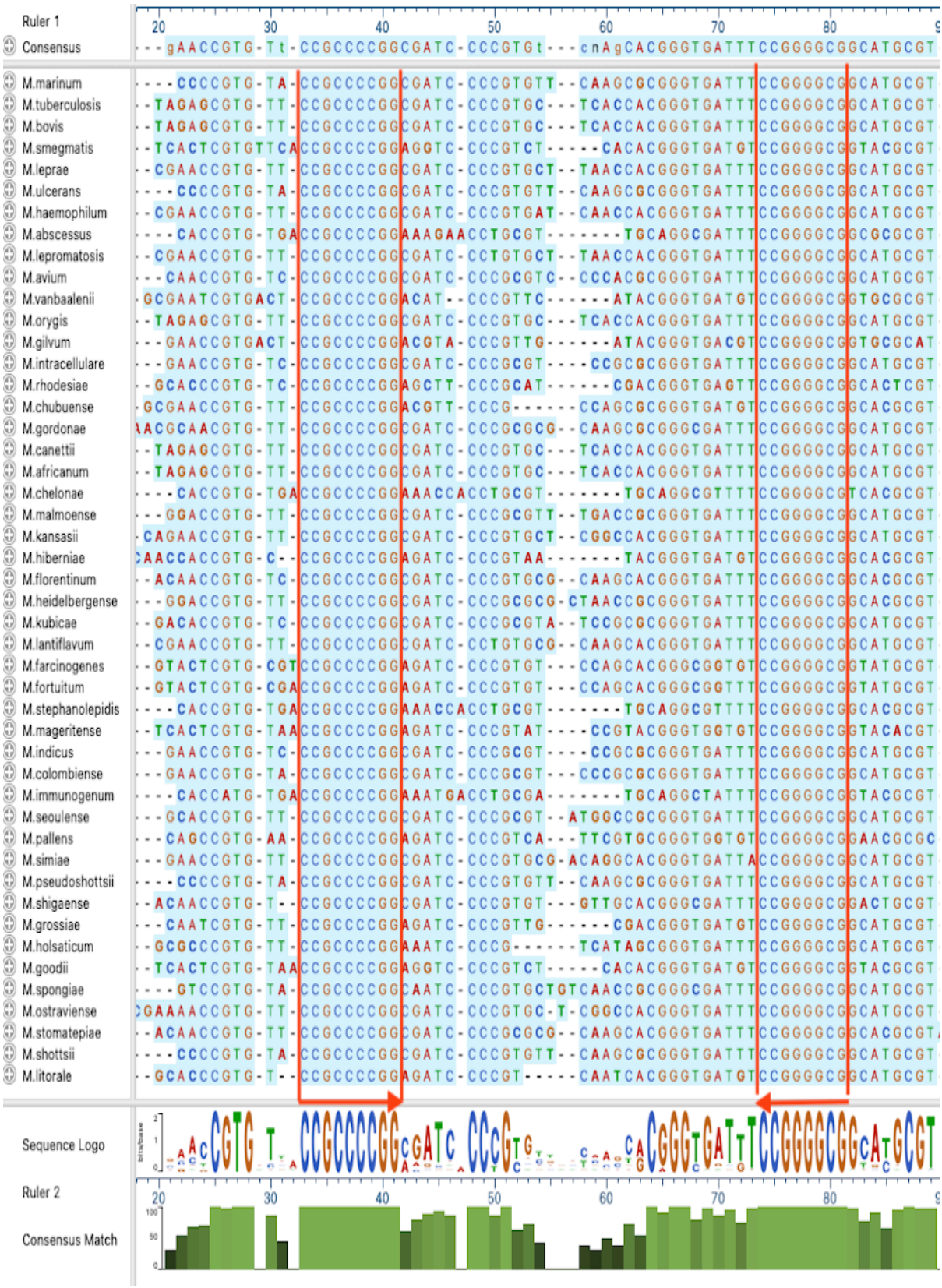


Figure 4.19 Alignment of upstream region of *cpn60.1* gene of 50 species of mycobacteria. The highly conserved inverted repeat sequence is marked with red lines.

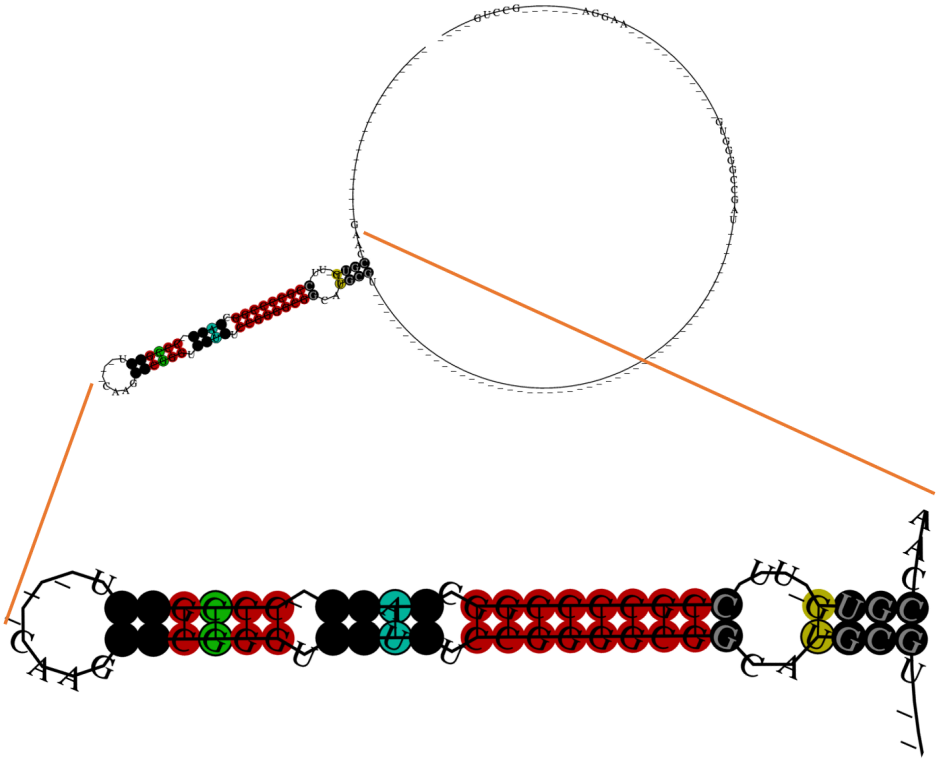


Figure 4.20 The MFE structure created using sequence alignment.
The MFE (Minimum Free Energy) structure (mRNA secondary structure) created using sequence alignment of the intergenic regions of *cpn10-cpn60.1* from 47 species of mycobacteria.

4.2.3 Bioinformatics analysis of HrcA repressor

The protein sequences of HrcA from eight species of actinobacteria, *B. subtilis* and *S. aureus* were obtained from NCBI database. The 8 species of actinobacteria included in the analysis are *M. tuberculosis*, *M. marinum, M. smegmatis, M. bovis, M. ulcerans, M. leprae, Streptomyces viridosporus* and *C. diphtheriae*. The multiple sequence alignment of all the ten HrcA protein sequences was performed using DNAstar Megalign pro software using default parameters of MUSCLE algorithm. The three conserved residues of *Bacillus* at the N-terminal potentially involved in the DNA binding i.e., S39G, A40G and R43D (Hitomi et al., 2003; Wiegert & Schumann, 2003) are indicated in red and blue arrows. Alanine and arginine were found to be conserved among all the 10 species, but serine was replaced with proline in *S. viridosporus* and *S. aureus* (Figure 4.21).

The pairwise amino acids sequence identity (%) and distance for the entire length of the HrcA homologues was also calculated using the same software. The lowest % identity calculated among the six mycobacterial species was 82.33, while the maximum distance was only 0.18. The HrcA of *M. marinum* and *M. tuberculosis* was found to be 93.06% identical with a calculated distance of 0.07. However, the amino acid identity among different genera was in the range of 24.92 – 59.31 %, which is surprisingly low. While the distance range was between 0.41 to 0.69 (Figure 4.22).

The multiple sequence alignment of the HrcA homologues was then used to construct the phylogenetic tree using maximum likelihood method. The sequences were bootstrapped 500 times. The phylogenetic tree is shown in Figure 4.23. The low G+C content species *B. subtilis* and *S. aureus* were clustered together while the actinobacteria species (high G+C) formed a separate group. The tree is consistent with the fact that *M. marinum* is known to be more closely related to *M. tuberculosis* than *M. smegmatis* as explained in Chapter 3 (section 3.3.2).

Overall, the analysis of HrcA homologues of the ten species used in the analysis shows that the sequence and structure of HrcA is surprisingly poorly conserved among different groups of bacteria.

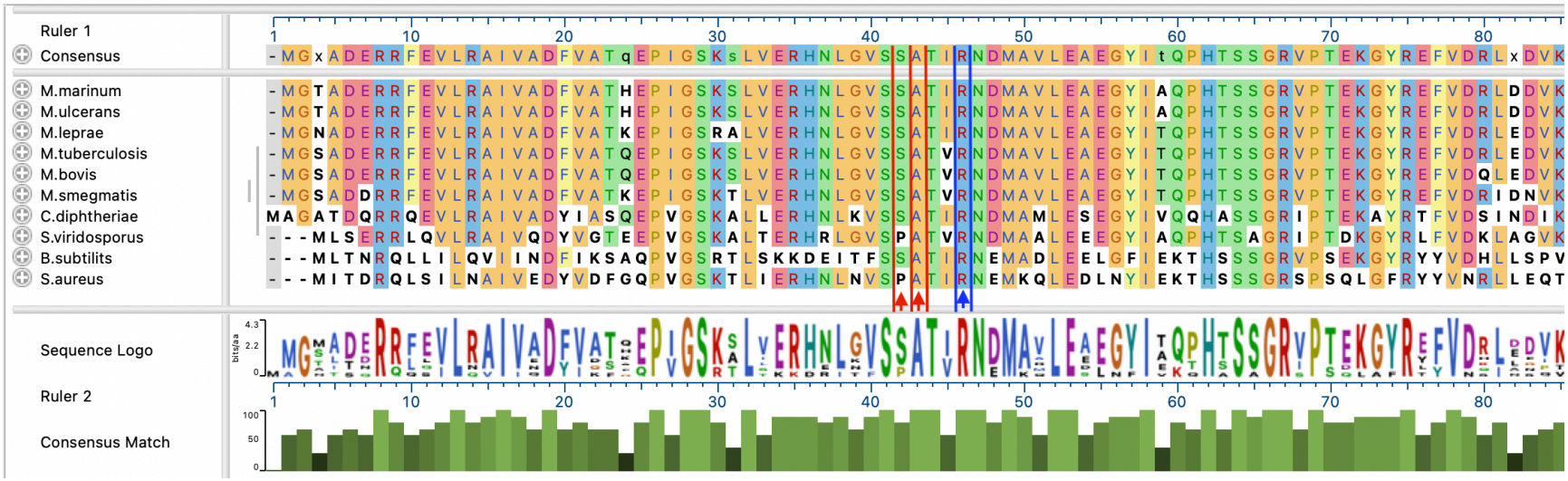


Figure 4.21 Multiple sequence alignment of HrcA homologues.
Multiple sequence alignment of HrcA homologues of ten different species. Only N-terminus residues are shown.

		A	B	C	D	E	F	G	H	I	J
A	M.marinum		0.01	0.10	0.07	0.07	0.15	0.43	0.43	0.73	0.73
B	M.ulcerans	99.37		0.11	0.08	0.08	0.15	0.44	0.43	0.73	0.73
C	M.leprae	89.59	88.96		0.09	0.10	0.18	0.44	0.43	0.73	0.73
D	M.tuberculosis	93.06	92.43	90.54		0.00	0.14	0.42	0.42	0.73	0.74
E	M.bovis	92.74	92.11	90.22	99.68		0.14	0.42	0.42	0.73	0.74
F	M.smegmatis	85.17	84.54	82.33	86.44	86.12		0.42	0.41	0.72	0.75
G	C.diphtheriae	57.10	56.47	56.47	57.73	57.73	58.04		0.49	0.73	0.74
H	S.viridosporus	57.41	56.78	57.41	58.36	58.36	59.31	51.42		0.74	0.73
I	B.subtilits	26.81	26.81	27.13	27.44	27.44	28.39	27.44	25.87		0.69
J	S.aureus	27.13	27.13	27.13	26.50	26.18	24.92	25.55	27.13	30.91	

Figure 4.22 Pairwise amino acid identity (%) and distance of HrcA from different species.
Pairwise amino acid identity (%) and distance for the full length of HrcA protein sequences from different species. The calculations were done using DNAstar Megalign pro software.

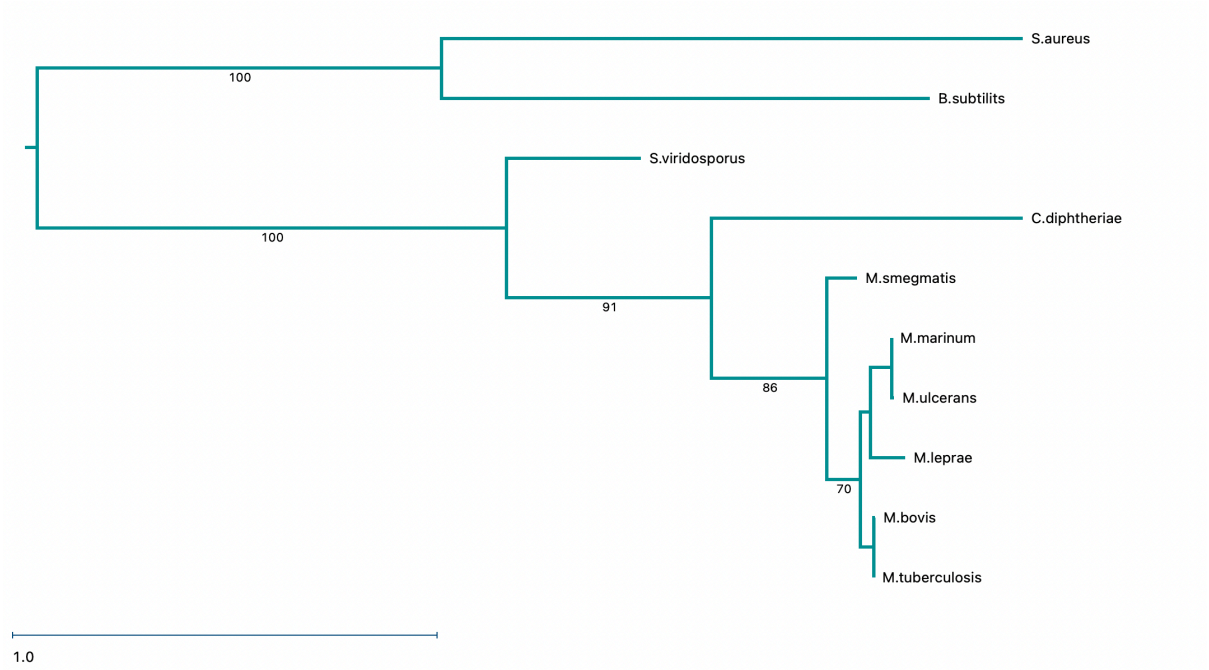


Figure 4.23 phylogenetic tree based on full length of HrcA protein sequences.
Maximum likelihood bootstrap phylogenetic tree based on full length of HrcA protein sequences of the ten different species. The numbers on nodes represent the % number of times the species were grouped together in 500 bootstraps.

4.3 Experimental plan

To study the promoter regions of *cpn* genes in *M. marinum*, different lengths of the upstream regions of *cpn10*, *cpn60.1* and *cpn60.2* were chosen to be investigated for promoter activity. For this purpose, a 9.5 kb shuttle vector pSD5B containing a promoter-less *lacZ* gene (encoding β-galactosidase) was used. The vector contains two origins of replication *oriE* (p15A *ori*) and *oriM* allowing it to propagate both in *E. coli* and mycobacteria. It has a kanamycin resistance cassette as a selection marker. The transcription terminator Ter-3 is present downstream of *oriM* to prevent readthrough transcription from *oriM*. Another strong transcriptional terminator Ter-4 is present downstream of *lacZ* gene to support the cloning of strong promoters (Gentz et al., 1981; Jain et al., 1997). Two restriction sites upstream to the *lacZ* gene, XbaI and SphI, were selected for the cloning of the fragments of interest (Figure 4.24). If the fragment cloned between these sites possess a promoter activity, then the expression of *lacZ* gene in a recombinant strain will result in the appearance of blue colour on X-Gal plates. The intensity of the blue colour can provide a qualitative measurement of the promoter activity of the cloned fragments. Besides, a quantitative β-galactosidase activity assay can provide a quantitative measurement of the *lacZ* transcripts which in turn is a measurement of the promoter activity.

As explained in section 4.1.5, CIRCE elements might play important role in the promoter activity of the regulatory regions of *cpn* operons. Three fragments of different lengths from the upstream region of *cpn10* were chosen to be cloned in the pSD5B vector between XbaI and SphI restriction sites; 1st fragment was 203 bps long starting from -1 position of the start codon of *cpn10* to the potential TSS (Figure 4.7), a 2nd fragment of 313 bps comprising the whole intergenic region of *cpn10* and *gcp* (the downstream gene), and a 3rd fragment of 369 bps. It was hypothesized that the 1st fragment will not possess any promoter activity because of the missing TSS and CIRCE elements while both the 2nd and 3rd fragments will show strong promoter activity.

To examine the promoter activity in the upstream region of *cpn60.1*, two fragments of 207 and 401 bps were chosen to be cloned, whereas one fragment of 309 bps was selected from the upstream region of *cpn60.2* (Figure 4.8). Based on the previous studies in *M. smegmatis* and *M. tuberculosis* (4.1.1 & 4.1.2), we hypothesized that the

upstream fragments of *cpn60.1* would not show any promoter activity while the upstream fragment of *cpn60.2* would demonstrate strong promoter activity. A schematic diagram of the experimental plan is shown in the Figure 4.25.

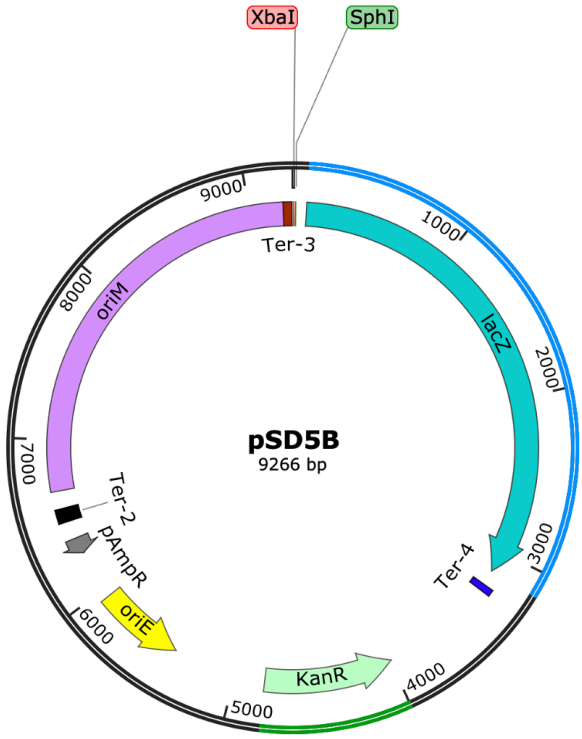


Figure 4.24 pSD5B shuttle vector.
A map of pSD5B shuttle vector showing *lacZ* gene, *kanR* cassette, *oriM*, *oriE* and restriction sites XbaI and SphI. Transcriptional terminators are also highlighted.

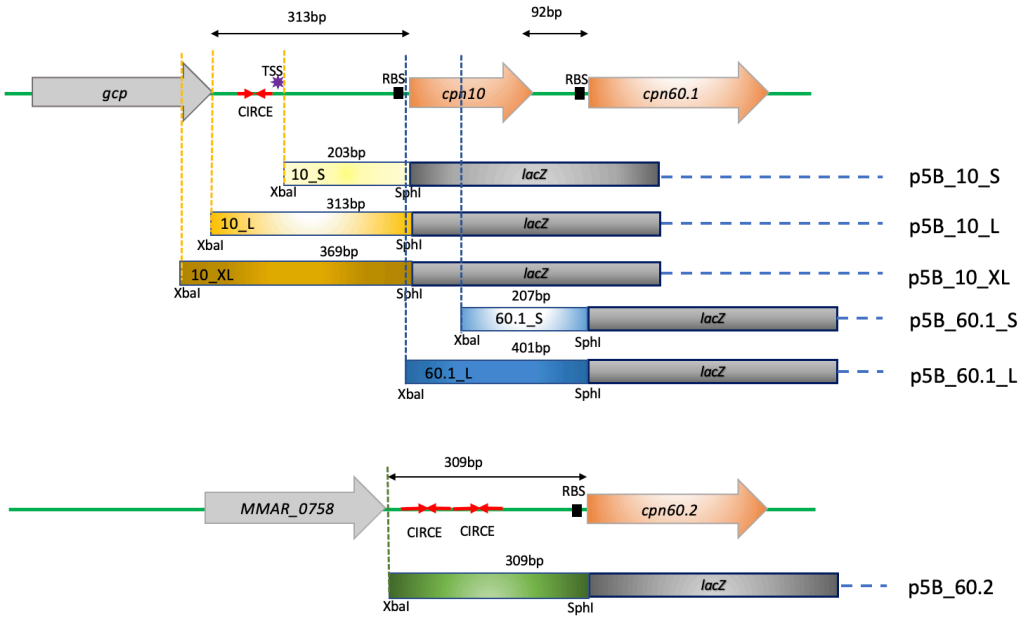


Figure 4.25 . Systematic plan of the fragments cloned into the plasmid pSD5B.
Three fragments of varying lengths from upstream region of *cpn10*, two fragments from that of *cpn60.1* while one fragment from upstream region of *cpn60.2* was cloned into SphI and XbaI sites of pSD5B. The coding regions are not scaled in this diagram.

4.4 Results

4.4.1 Cloning of fragments into pSD5B

The genomic DNA of *M. marinum* 1218R was extracted using protocol given in chapter 2 (2.4.5). Different lengths of the upstream fragments of *cpn* genes (Figure 4.25) were amplified from the genomic DNA of *M. marinum* 1218R using PCR amplification. Oligonucleotide primers used for the amplification of the fragments were synthesized by Sigma Aldrich and are listed in Table 2.3 (Chapter 2). All the forward primers contained an XbaI restriction site while all the reverse primers had SphI restriction site.

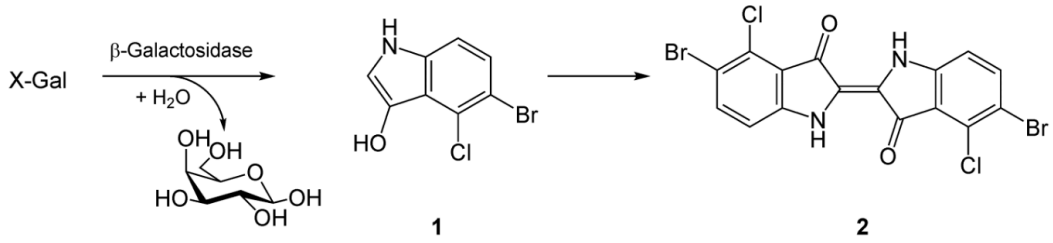
4.4.1.1 Direct cloning of PCR fragments into pSD5B

For single step cloning, the PCR amplification reaction was set up using MyTaq DNA polymerase. 2 ml from each of the PCR reaction was loaded on 1% agarose gel to confirm the accuracy of the amplified products. The PCR reactions were then purified using QIAquick PCR Purification Kit. The pSD5B plasmid and the PCR amplified fragments were digested with the restriction enzymes XbaI and SphI. Following gel purification, the fragments were attempted to ligate directly into the XbaI and SphI sites of the plasmid. However, this attempt was not successful. It was then decided to use the two-step cloning strategy i.e., cloning of the PCR fragments into pUC18 and then from pUC18 into the promoter probe vector pSD5B.

4.4.1.2 Two step cloning strategy using pUC18

pUC18 is a high copy number plasmid containing *lacZα* gene which encodes the N-terminal part (α-peptide) of the β-galactosidase enzyme under control of *lac* operon. Multiple cloning sites (MCS) are present within the 5' terminal part of the *lacZ*α. TOP10 and DH5α are *E. coli* (lacZ)ΔM15 laboratory strains (a strain having N-terminal part of *lacZ* gene deleted) which encode only for the ω-fragment of β-galactosidase. pUC18 normally forms blue colonies on an agar plate containing X-Gal and IPTG when expressed in TOP10 or DH5α because of the formation of a functional copy of the β-galactosidase by the combination of α-peptide encoded by plasmid and ω-fragment encoded by the chromosomal copy of the lacZ gene (gene coding for β-galactosidase). Cloning a fragment into MCS of pUC18 disrupts the α-peptide, so a non-functional β-

galactosidase enzyme is produced resulting in white colonies on X-Gal agar plates. X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) is a colourless organic compound used in molecular biology to test the presence of β-galactosidase enzyme. When β-galactosidase hydrolyse the X-Gal, 5-bromo-4-chloro-indoxyl is produced which quickly dimerize to produce an insoluble intensively blue colour compound called 5,5'-dibromo-4,4'-dichloro-indigo. Therefore, the blue colonies on X-Gal plate indicate that the cells produce a functional copy of β-galactosidase enzyme (Horwitz et al., 1964).



The reaction above shows the breakdown of X-Gal by β-galactosidase and production of 5-bromo-4-chloro-indoxyl (1). Dimerization of (1) produces 5,5'-dibromo-4,4'-dichloro-indigo (2) which is an insoluble blue colour compound.

Step-1: Cloning into pUC18

Taking advantage of blue/white screening, it was decided to clone all the fragments through blunt end ligation into SmaI site present in the MCS of pUC18. MyTaq DNA polymerase cannot be used in PCR amplification for blunt end ligation because it produces fragments with A˙ overhangs at 3'end, so pfu DNA polymerase was used to amplify all the fragments from genomic DNA of *M. marinum* 1218R. PCR reaction set up and PCR parameters used for amplification are described in the tables 2.5 & 2.6. pUC18 plasmid was digested with SmaI restriction enzyme for one hour at 25°C and gel purified. The PCR amplified fragments were purified and ligated into SmaI restriction site of pUC18. Along with LB agar, X-Gal (for blue/white selection), IPTG and ampicillin (selection marker for pUC18) were added in the transformation plates. Successfully cloned pUC18 derivatives were screened by colony PCR of white colonies. Same forward and reverse primers were used for PCR screening which were used for amplification of the fragments from genomic DNA. All fragments were cloned

successfully into pUC18 and confirmed by sequenced using standard primers for pUC18 (i.e., M13_F and M13_R).

Step-2: Cloning from pUC18 to pSD5B

All the fragments cloned into pUC18 were confirmed by sequencing. The pUC18 derivatives were digested with SphI and XbaI. Followed by gel purification of the digested fragments, purified fragments were then ligated into SphI and XbaI sites of pSD5B shuttle vector.

Two issues were faced during cloning of fragments from pUC18 to pSD5B.

1) XbaI methylation problem

The activity of XbaI is blocked by overlying dam methylation. XbaI restriction site followed by TC or headed by GA is methylated at GATC site by Dam methylase in most *E. coli* strains. In second fragment of *cpn10*-L (313 bps), there was a TC site next to XbaI site, therefore, XbaI was not cutting during restriction digestion of pUC18 derivative containing this fragment. To solve this problem pUC18 derivative plasmid was propagated in JM110 (which is a dam deficient mutant (*dam-/dcm-* mutant) of *E. coli*) and purified for restriction digestion.

2) Presence of intrinsic SphI site in the upstream region of cpn60.1

There was an intrinsic SphI site 31 bps upstream to *cpn60.1* start codon (Fig 4.23). This site was not noticed before cloning of fragments into pUC18. For screening of successfully cloned pUC18 derivatives of the two *cpn60.1* upstream fragments, a new reverse primer was designed (named cpn60.1_S/L_R). Both the small and large fragments upstream of *cpn60.1* were digested from pUC18 and successfully cloned into pSD5B but with missing 31 bps upstream to *cpn60.1* start codon. The missing 31 bps (called Sph fragment) in the plasmids p5B_60.1_S and p5B_60.1_L were then cloned in the SphI site. Both the forward and reverse strands of 31 bps oligos (given in chapter 2, section 2.3) were prepared from source biosciences (https://www.sourcebioscience.com).

Both the forward and reverse oligos of SphI fragment were phosphorylated separately with T4PNK enzyme according to the reaction mixture parameters given in Table 2.18. The reaction mixtures were incubated at 37°C for 1 hour followed by heat inactivation of T4PNK for 20 minutes.

For annealing the oligos, 10 µL of each fragment was mixed in 80 µl of Tris Buffer Saline (TBS). The mixture was heated at 95°C for 3 minutes and then allowed to cool gradually to room temperature (~25°C). The phosphorylated and annealed oligos were then ligated into SphI site of both plasmids, p5B_60.1_S and p5B_60.1_L, using the standard ligation reaction conditions with a vector to fragment ratio of 1:10.

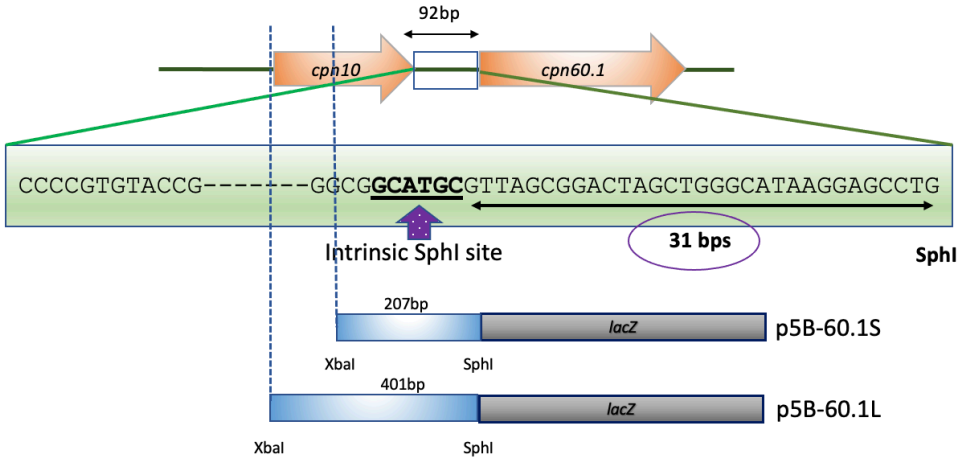


Figure 4.26 Intrinsic SphI site 31 bps upstream to *cpn60.1* start codon.
These 31 bps were then cloned separately in p5B_60.1_S and p5B_60.1_L plasmid constructs.

All the fragments were successfully cloned into pSD5B and confirmed by sequencing using primer pSD5B_R; the names given to pSD5B derivative plasmids are listed in Table 2.2. All the plasmids were introduced in *E. coli* TOP10 strain by transformation. Electrocompetent cells of *M. smegmatis* MC²155, *M. marinum* 1218R (W.T) and *M. marinum* 1218R Δ*cpn60.1* were then electroporated with all the promoter probe plasmids as described in chapter 2 (Table 2.2). All the recombinant strains were then used for qualitative and quantitative analysis of promoter activity in the cloned fragments. *M. smegmatis* and *E. coli* were used to investigate if the promoters of *M. mairnum* work efficiently in these species or not.

4.4.2 Qualitative X-Gal assay

For qualitative measurement of promoter activity, the recombinant strains were grown using appropriate media and culture conditions till the mid log phase. The OD_{600} of all the cultures was measured and used to normalize the cultures. Serial dilutions were made, and spotting was done on X-Gal LB agar plate in case of *E. coli* TOP10 recombinant strains and on X-Gal 7H10 agar plates in case of mycobacterial recombinant strains. All the plates contained kanamycin as a selection marker. However, the recombinant strains of *M. marinum* 1218R Δcpn60.1 (KO) were also spotted on plate containing both hygromycin and kanamycin (containing X-Gal) to confirm the *cpn60.1* deletion strain. All the *M. marinum* KO recombinant strains showed colonies on the hygromycin plate (Figure 4.27 C). The detailed protocols for the X-Gal plates preparation and spotting assay are given in chapter 2 (section 2.5).

4.4.2.1 Qualitative assay with *M. marinum* strains

The Plates spotted with *M. marinum* strains were incubated at 30°C for seven days. Both in the W.T and KO strains, out of the three fragments upstream to *cpn10*, the two fragments 10_L (313 bps) & 10_XL (369 bps) showed a strong promoter activity while the small fragment (10_S, 203 bps) did not show any promoter activity. The small fragment is the upstream region from -1 position of the start codon of *cpn10* to +1 position of the potential TSS (Figure 4.7). The absence of promoter activity in this region while a strong promoter activity in the fragment 10_L indicated that the promoter activity is present in the region around CIRCE element. This reinforces the idea of HrcA/CIRCE regulation mechanism for the promoter regions of mycobacterial chaperonin. Neither the shorter nor the longer fragment upstream of *cpn60.1* showed any promoter activity. However, a strong promoter activity was observed in the fragment of the *cpn60.2* upstream region (Figure 4.27 A & B).

4.4.2.2 Qualitative assay with *M. smegmatis* and *E. coli*

The *E. coli* TOP10 strains harbouring the promoter probe plasmids did not show promoter activity in any of the fragments except a very weak signal of blue colour in the strain harbouring p5B_60.2 plasmid (Figure 4.28, plate A). This is consistent with the fact that mycobacterial promoters do not usually function well in *E. coli*.

In *M. smegmatis* MC²155, the two large fragments upstream of *cpn10* (10_L & 10_XL) showed promoter activity by developing the blue colour on X_Gal plate while the small fragment did not develop blue colour indicating no promoter activity in the 10_S fragment. Also, neither the short nor the long fragment upstream of *cpn60.1* showed promoter activity in *M. smegmatis* but the fragment upstream of *cpn60.2* strong showed promoter activity. The qualitative X-Gal assay in *M. smegmatis* presented results comparable to those of *M. marinum*. However, the intensity of the blue colour developed in *M. smegmatis* strains was noticed to be less strong than the colour produced in *M. marinum* strains (Figure 4.28, plate B).

The Figure 4.27 shows that the strains harbouring vector alone (negative control) developed a pale blue colour on X-Gal plates whereas some other strains harbouring derivative plasmids 2, 5 and 6 (plate A) developed white colour. The possible reason for the pale blue colour in the colonies of vector alone strains could be due to the reason that Ter-3 site might not completely blocked the transcription from the previous promoter (Figure 4.24) and the insertion of a fragment (promoter-less) caused the complete blockage of the transcription.

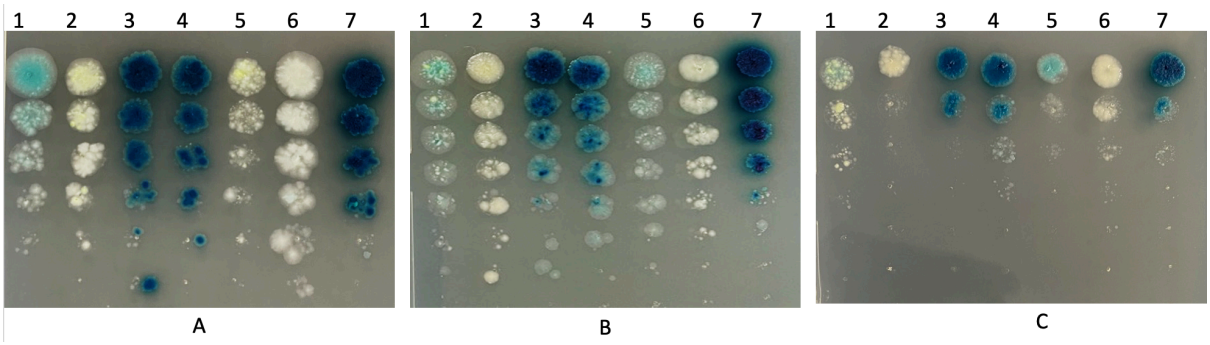


Figure 4.27 Qualitative assay with *M. marinum* strains.
7H10 agar plates supplemented with X-Gal and kanamycin (selection marker for pSD5B and derivative plasmids). Plate C also contains hygromycin (selection marker for KO strain *M. marinum* 1218R-Δcpn60.1). Plate A is spotted with the recombinant strains of *M. marinum* 1218R, plate B and C are spotted with *M. marinum* 1218R-Δcpn60.1 recombinant strains. Recombinant strains contain plasmids 1 (pSD5B – negative control), 2 (p5B_10_S), 3 (P5B_10_L), 4 (P5B_10_XL), 5 (P5B_60.1_S), 6 (P5B_60.1_L) and 7 (P5B_60.2). Pictures were taken at the seventh day of incubation at 30°C. Less growth on plate C than plate B could be due to the slower growth caused by double selection markers (hygromycin + kanamycin).

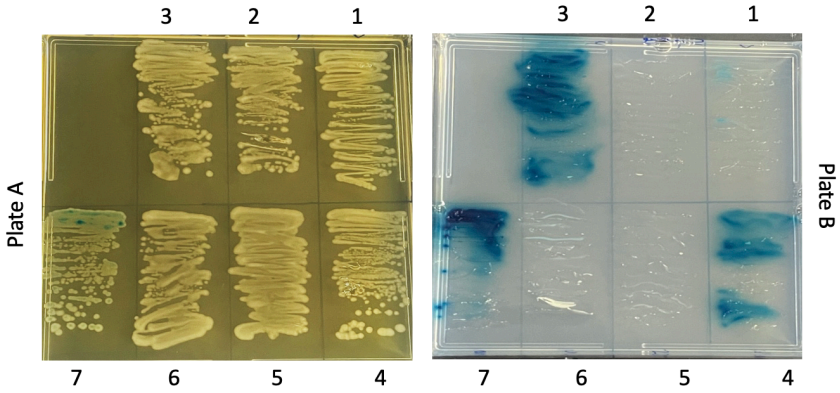


Figure 4.28 Qualitative assay with *M. smegmatis* and *E. coli* strains.
Plate A, LB agar plate containing X-Gal and Kanamycin. Plate streaked with *E. coli* TOP10 strains harbouring the negative control vector pSD5B and all pSD5B derivative plasmids. Plate B, 7H10 agar plate supplemented with X-Gal and kanamycin, streaked with *M. smegmatis* MC^2155 harbouring the negative control vector pSD5B and all pSD5B derivative plasmids. Recombinant strains contain plasmids 1 (pSD5B), 2 (p5B_10_S), 3 (P5B_10_L), 4 (P5B_10_XL), 5 (P5B_60.1_S), 6 (P5B_60.1_L) and 7 (P5B_60.2). For plate A, picture was taken after 24 hours of incubation at 37°C while plate B was incubated at 37°C for three days.

4.4.3 Quantitative β-galactosidase activity assay

The qualitative β-galactosidase assay using X-Gal plates only showed the presence or absence of promoter activity in the upstream fragments of *cpn* genes; the assay does not provide information about the relative strength of the promoter elements. To estimate the strength of promotor activity quantitative β-galactosidase assay was used. In this procedure a synthetic compound ONPG (o-nitrophenyl-β-D-galactopyranoside) is used which is cleaved by β-galactosidase yielding galactose and a yellow colour compound called o-nitrophenol. In the presence of excess of ONPG, the concentration of β-galactosidase in the reaction mixture can be estimated by the production of o-nitrophenol. Greater the amount of o-nitrophenol per unit time, greater the β-galactosidase concentration which is measured in Miller units (Smale, 2010). The concentration of β-galactosidase is in turn directly proportional to the strength of the *lacZ*-expressing promoter. The detailed protocol and the formula used for the calculations of the Miller units is given in chapter 2 (Section 2.5.2).

4.4.3.1 Quantitative β-galactosidase assay in *M. marinum* W.T and KO strains at normal growth conditions

In the first attempt, all the *M. marinum* 1218R W.T recombinant strains harbouring plasmids pSD5B, p5B_10_S, P5B_10_L, P5B_10_XL, P5B_60.1_S, P5B_60.1_L and P5B_60.2 (Figure 4.25) were used for quantitative β-galactosidase assay (or *lacZ* activity assay) at normal growth and temperature condition (30°C). For each strain, three biological replicates of secondary cultures were grown in 7H9 liquid media in the presence of kanamycin at 30°C till the OD_{600} between 0.8 to 1.0. All the cultures were normalized, and 100 µl of each sample was used to perform the β-galactosidase assay. Strongest promoter activity was found in the fragment upstream to *cpn60.2* (*cpn60.2* promoter). The short fragment upstream to *cpn10* did not show any promoter activity which again confirms the important role of CIRCE element in the promoter activity. The two other fragments (L and XL) upstream to the *cpn10* showed nearly equal levels of promoter activity (*cpn10* promoter). No promoter activity was recorded in any of the two fragments upstream to *cpn60.1*. In addition, the promoter activity of *cpn60.2* promoter was found to be more than double the amount of the promoter activity of *cpn10* promoter (Figure 4.29).

The assay was repeated with *M. marinum* 1218R-Δ*cpn60.1* recombinant strains harbouring all the promoter probe plasmids (Figure 4.22). W.T strains were included in the experiment for comparison. The purpose of performing the *lacZ* assay using the *cpn60.1* KO strain was to investigate if the absence of Cpn60.1 causes any increase or decrease in the promoter activity of *cpn* promoters. All the fragments showed similar results as was observed in the W.T strains; highest promoter activity was calculated in the fragment upstream to *cpn60.2* followed by the fragment upstream to *cpn10* whereas no other fragment showed any promoter activity (Figure 4.30). However, the assay showed that the activity of the *cpn60.2* promoter was increased around 25% in the KO strain compared to the W.T strain. And the activity of *cpn10* promoter (in both the L and XL fragments) was also observed to be increased by nearly 35% in the KO strain compared to its activity in the W.T strain (Figure 4.30). The results show that the absence of Cpn60.1 might cause an increase in the expression of the second chaperonin gene *cpn60.2* and the co-chaperonin *cpn10*.

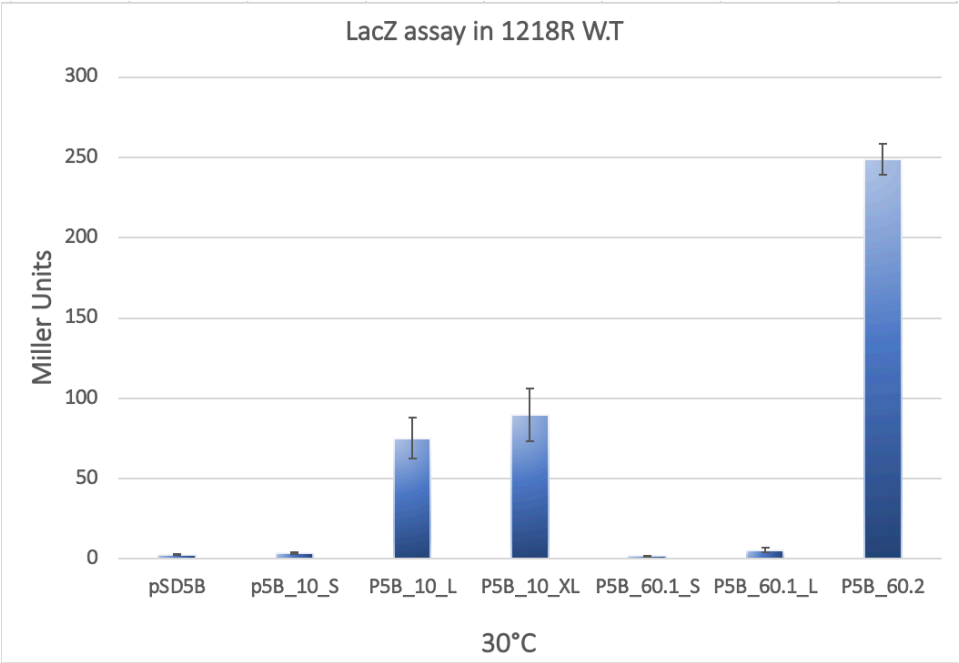


Figure 4.29 β-galactosidase assay in *M. marinum* W.T strain.
β-galactosidase assay in *M. marinum* W.T strain harbouring plasmids pSD5B (negative control), p5B_10_S, P5B_10_L, P5B_10_XL, P5B_60.1_S, P5B_60.1_L and P5B_60.2. The assay was performed using the cultures grown at 30°C under normal growth and antibiotic conditions.

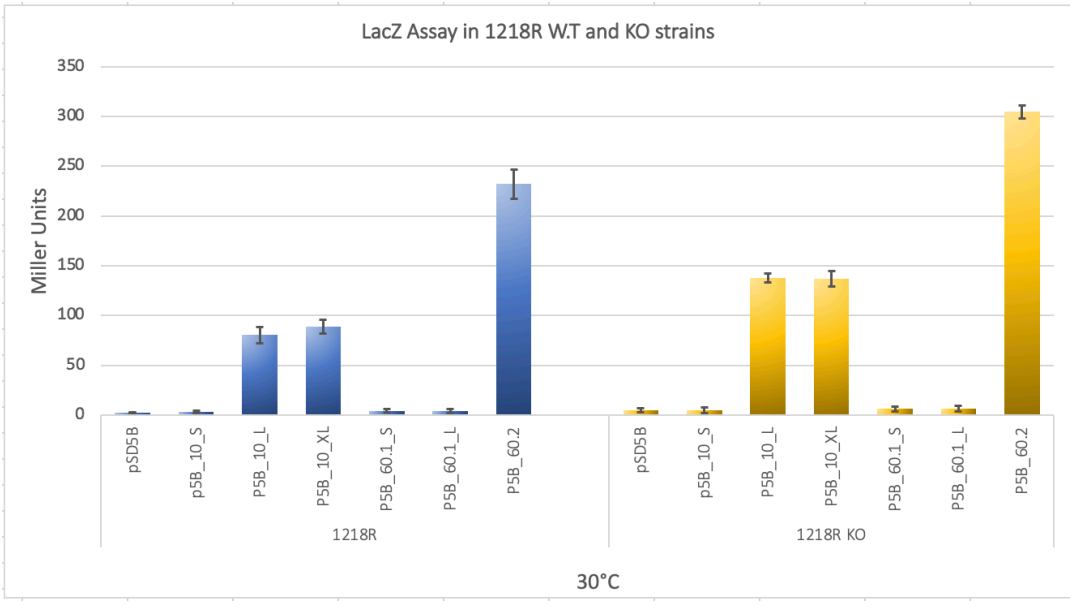


Figure 4.30 β-galactosidase assay in *M. marinum* W.T and KO strains.
β-galactosidase assay in *M. marinum* W.T and KO strains harbouring plasmids pSD5B (negative control), p5B_10_S, P5B_10_L, P5B_10_XL, P5B_60.1_S, P5B_60.1_L and P5B_60.2. All the cultures were grown at 30°C under normal growth and antibiotic conditions.

4.4.3.2 Quantitative β-galactosidase assay in *M. marinum* W.T and KO strains at temperatures 5°C lower and higher than the optimum temperature i.e., 30°C.

The optimum temperature range for *M. marinum* is from 25°C to 30°C (Clark & Shepard, 1963). To investigate if the change in temperature have any effect on the expression levels of the *cpn* genes, the promoter activity was calculated at 25°C, 30°C and 35°C (heat shock). For this purpose, β-galactosidase activity assay was performed in both the W.T and KO 1218R strains. The triplicate samples of the secondary growth cultures were grown initially at 25°C till the $O.D_{600}$ reached to 0.8 to 1.0. Each sample was then divided in three equal aliquots (5ml each). One set of aliquots was heat shocked at 35°C in a shaking incubator for 1 hour. The second set of aliquots was grown at 30°C in the shaking incubator for 1 hour. The 3rd set was kept at 25°C. All the samples were then used to perform the β-galactosidase assay.

The activity of *cpn60.2* promoter was found to be increased with change in temperature from 25°C to 30°C and 35°C both in the W.T and KO strains (Figure 4.31). A temperature shift from 25°C to 30°C increased the *cpn60.2* promoter activity nearly 50% in the W.T strain and nearly 25% in the KO strain whereas a further 5°C increase in temperature only caused slight increase in the promoter activity i.e., nearly 15% in both the W.T and KO strains. On the other hand, *cpn10* promoter showed a small increase of nearly 30% in the W.T strain at a temperature shift from 25°C to 30°C although a further increase to 35°C did not increased the promoter activity any further. Similar results were found in the KO strain, a temperature shift from 25°C to 30°C caused nearly 35% increase in the *cpn10* promoter activity while a further temperature shift of 5°C did not show a significant increase. The results were similar both in case of L and XL fragments of *cpn10* upstream region. The assay also demonstrated that at all the three temperatures the promoter activity of both the *cpn60.2* and *cpn10* promoters was higher in the KO strain compared to the W.T strain (Figure 4.31). No promoter activity was observed in the short fragment of *cpn10* upstream and in both the fragments of *cpn60.1* upstream, consistent with the results of section 4.4.5.1.

4.4.3.3 Quantitative β-galactosidase assay in *M. marinum* W.T and KO strains at pH stress

The assay was also done at two pH stress conditions (pH 4: acidic stress and pH 10: alkaline stress). Both stress conditions caused an overall 20-30% increase in the activity of *cpn60.2* promoter both in W.T and KO strains. Slight (insignificant) increase in the *cpn10* promoter activity was observed at both pH stress conditions (Figure 4.32).

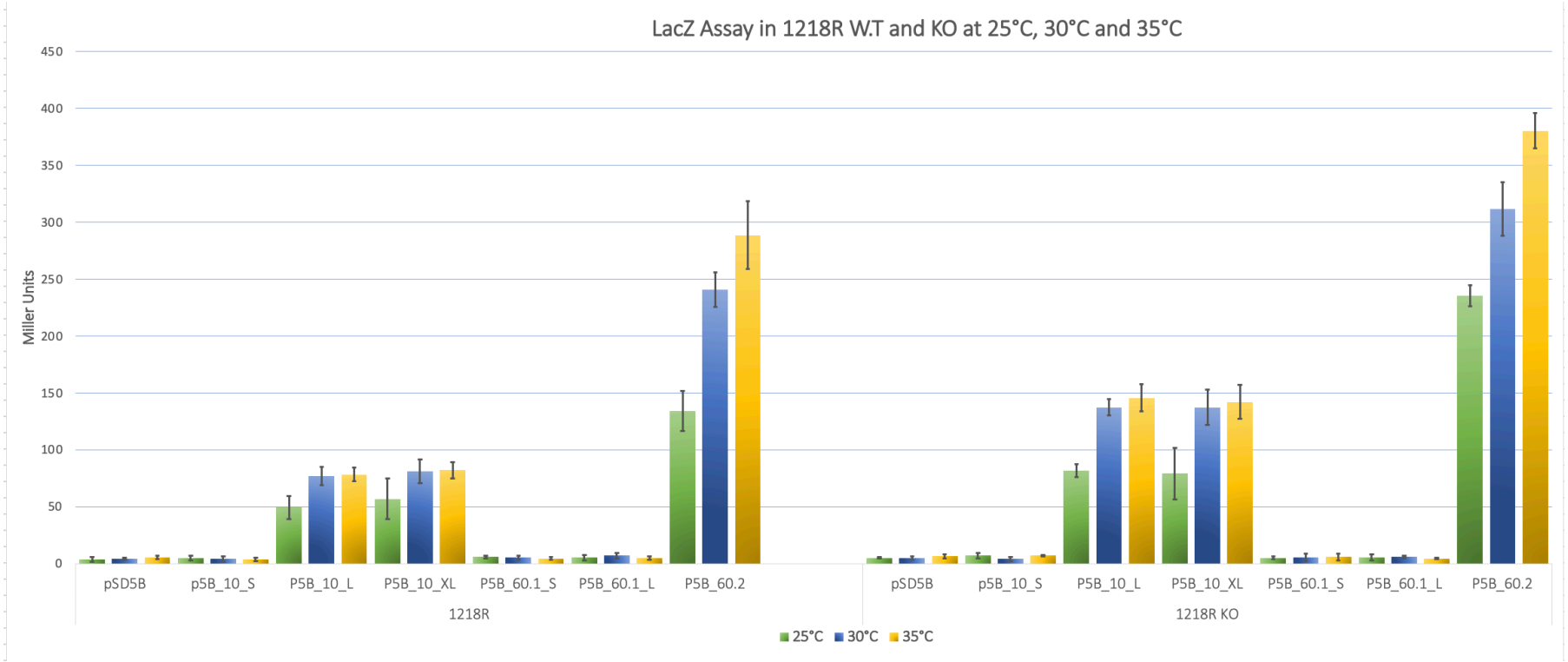


Figure 4.31 β-galactosidase assay in *M. marinum* W.T and KO strains at different temperatures.
β-galactosidase assay in *M. marinum* W.T and KO strains harbouring plasmids pSD5B (negative control), p5B_10_S, P5B_10_L, P5B_10_XL, P5B_60.1_S, P5B_60.1_L and P5B_60.2 at three different temperatures (25°C (lowest optimum), 30°C (highest optimum) and 35°C (heat-shock)).

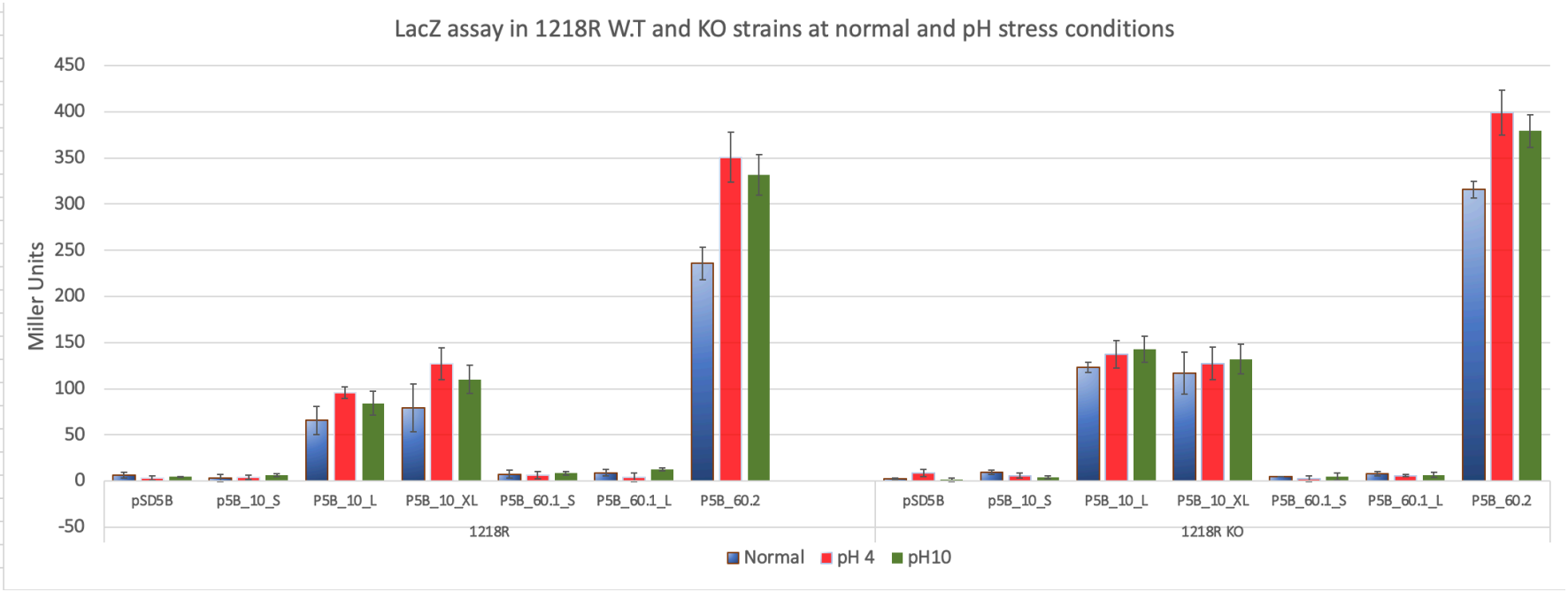


Figure 4.32 β-galactosidase assay in *M. marinum* W.T and KO strains at different pH conditions.
β-galactosidase assay in *M. marinum* W.T and KO strains harbouring plasmids pSD5B (negative control), p5B_10_S, P5B_10_L, P5B_10_XL, P5B_60.1_S, P5B_60.1_L and P5B_60.2 at three conditions; normal growth condition, pH stress at pH 4 (acidic) and pH 10 (alkaline).

4.4.3.4Quantitative β-galactosidase assay in *M. smegmatis*

To investigate if the M. marinum promoters work efficiently in *M. smegmatis*. *M. smegmatis* MC²155 strain harbouring *M. marinum* promoter probe plasmids pSD5B (negative control), p5B_10_S, P5B_10_L, P5B_10_XL, P5B_60.1_S, P5B_60.1_L and P5B_60.2 was used for performing the β-galactosidase assay. Both the *cpn60.2* and *cpn10* promoters showed reasonable promoter activity but the values Miller units calculated for this experiment were much smaller than those calculated in β-galactosidase assay done in *M. marinum*. Both promoters showed only half of the average Miller unit values obtained for assays with *M. marinum* which shows that although *M. marinum* promoters are functional in *M. smegmatis*, but the activity is much less than in the original organism they belong to (Figure 4.33).

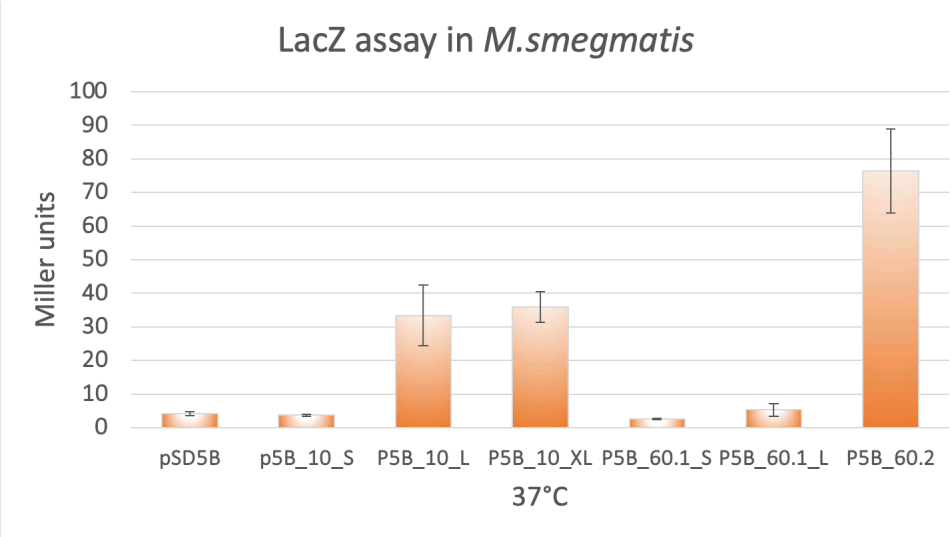


Figure 4.33 β-galactosidase assay in *M. smegmatis*.
β-galactosidase assay in *M. smegmatis* harbouring *M. marinum* promoter probe plasmids pSD5B (negative control), p5B_10_S, P5B_10_L, P5B_10_XL, P5B_60.1_S, P5B_60.1_L and P5B_60.2.

4.5 Discussion

Promoter probe analysis demonstrated strong promoter activity in the upstream regions of both the Mm*cpn10* and Mm*cpn60.2* genes. The absence of promoter activity in the short fragment upstream to Mm*cpn10* could be due to the missing TSS in that fragment. The short fragment (p5B_10_S) also did not have the CIRCE element which is thought to be involved in regulation of the promoter activity. Under normal conditions of growth and temperature, the promoter activity of fragment upstream to Mm*cpn60.2* was found to be more than 2-fold higher than that of Mm*cpn10*.

The results also revealed that Mm*cpn60.2* is likely to be the more highly expressed chaperonin gene in *M. marinum* both under normal and stressed conditions followed by *cpn10* because *cpn60.2* promoter was found to have the strongest promoter activity under normal, heat-shock and all the stressed conditions tested in this part of the study. A 5°C shift in temperature (from 25°C to 30°C) caused significant (nearly double) increase in the activity of *cpn60.2* promoter which increased slightly more with another 5°C temperature shift (30°C to 35°C) suggesting that increase in temperature increases the expression of *cpn60.2* in *M. marinum.* In case of acid and alkaline stress conditions, the activity of Mm*cpn60.2* promoter was also increased by 20-30%. These findings indicate that *M. marinum* Cpn60.2 might be the major housekeeping chaperonin which is induced at stressed conditions. The idea is consistent with the previous studies of chaperonin genes of *M. tuberculosis* and *M. smegmatis*.

Mm*cpn10* promoter showed increase (30-35%) in the activity at first 5°C of temperature shift (from 25°C to 30°C) but further increase in temperature (from 30°C to 35°C) did not cause any further induction of the *cpn10* promoter. This suggests that increase in temperature also induces the *M. marinum cpn10* promoter, but the induction is less strong than the Mm*cpn60.2* promoter. At acid and alkaline stress, more increase was observed in Mm*cpn60.2* promoter activity than Mm*cpn10*. *cpn10* promoter was constitutive or increased only slightly (insignificant).

In addition, promoter probe study did not provide any information about the expression levels of Cpn60.1 but it was observed that no promoter activity was found in the intergenic region of Mm*cpn10* and Mm*cpn60.1*. These findings confirm that *M. marinum cpn60.1* exists in a single operon with *cpn10,* and that some unknown

mechanism controls the expression levels of both genes. It was also observed that the *M. marinum* promoters function less well in *M. smegmatis* whereas the promoter activity was not observed or was negligible (in case of Mm*cpn60.2* promoter) in *E. coli*.

Moreover, a very interesting findings of this part of the study was the increase in the activity of Mm*cpn10* and Mm*cpn60.2* promoters (nearly 35% and 25% respectively) in the KO *M. marinum* strain. This shows that the absence of Cpn60.1 caused an increase in the expression of the other chaperonin homologue (Cpn60.2) and the co-chaperonin (Cpn10). This could be because MmCpn60.1 might preforms some chaperone activity in its natural cell environment and in the absence of MmCpn60.1 the role is compensated by the overproduction of MmCpn60.2.

To further study this it would be interesting to look at the effect of deletion of Cpn60.1 on expression of other genes to see (a) whether other chaperones are affected by the KO and (b) what other genes might be changed as this could potentially give us further insights into the role of MmCpn60.1. This could have been done using more promoter probes, but limitation here is that you have to choose which genes to look at. So, we decided to take a whole transcriptome approach.

We investigated the expression pattern of *cpn* genes along with the whole transcriptome profiling of *M. marinum* 1218R using the W.T and KO and complemented strains. In the next chapter the detailed study of RNA-seq analysis is used to examine the hypothesis further.

5 Transcriptome analysis of *M. marinum* based on RNA-seq

5.1 Introduction

As described in the chapter 1 (section 1.4.3, also see chapter 6, section 6.1), the complementation studies of chaperonins of *M. tuberculosis* and *M. smegmatis* consistently showed that Cpn60.2 of both the species alone could complement the loss of GroEL in *E. coli* whereas Cpn60.1 failed to complement under similar conditions. Similar results were found with the chaperonins of *M. marinum* which will be discussed in next chapter. Briefly, the results are consistent with the hypothesis that Cpn60.2 of *M. marinum* functions as a housekeeping chaperonin while Cpn60.1 might not function as a chaperone in *M. marinum*. Previous studies have shown the similar findings for other mycobacterial species such as *M. tuberculosis* and *M. smegmatis* (see chapter 6, section 6.1). In addition, many studies on *M. tuberculosis* and *M. smegmatis* demonstrated the differential expression of *cpn60.1* and *cpn10* as discussed in the introduction section of chapter 4.

To get insight into the role of Cpn60.1 in *M. marinum* and to investigate the impact of deletion of *cpn60.1* on overall gene expression we therefore decided to study the transcriptome (a full set of all the mRNAs in an organism) profile of the *M. marinum* 1218R W.T, the *cpn60.1* knockout *(Δcpn60.1*) and complemented strains using high-throughput mRNA sequencing (RNA-seq) approach. Our hypothesis was that changes in transcriptome profile in the deletion mutant might give us some insights into the cellular role of *cpn60.1*.

For better understanding of RNA sequencing technique, the section below will describe a brief comparison of different methods available for gene expression profiling and an overview of the RNA-seq method.

5.1.1 Comparison of gene expression analysis methods

Various methods are available for gene expression studies. Among low-throughput methods northern blotting and qRT-PCR (quantitative real time-PCR) are generally used which allow the study of one to few genes. In northern blotting method, a specific RNA molecule is detected by binding of radioactive or fluorescent labelled probes to RNA whereas in qRT-PCR sequence specific primers bind to the target cDNA to amplify

a specific region of the transcript and the quantity is measured by fluorescent reporter dyes (Croucher & Thomson, 2010).

In the beginning of 21st century, hybridization-based microarray and tiling array methods were used to study the expression of large number of genes and expression of the entire genome respectively. Both techniques are based on the same principle of hybridization of cDNA to probes (Marioni et al., 2008; Moody, 2001). Tiling array is emphatically expensive because it requires large number of probes while microarray is relatively inexpensive. But the hybridization-based technology has significant limitations such as low dynamic range, high noise, and artifacts due to cross-hybridization. Because of which these techniques are not capable of quantifying the levels of expression of least and/or highly expressed genes accurately. Other limitations are low accuracy in covering the full transcripts and requirement of difficult and complicated statistical analysis methods to compare the expression data obtained from different experiments. In addition, array technology cannot be used for the study of expression of those genes/genomes for which the sequence is not already known (Kukurba & Montgomery, 2015; Sorek & Cossart, 2010).

To overcome the technical limitations of hybridization-based methods, higher throughput tag-based methods such as cap analysis gene expression (CAGE) and serial analysis of gene expression (SAGE) were developed for studying the transcriptome profile. These methods offer clear advantage over array-based methods in providing more accurate quantification of genes expression. However, many disadvantages like labour intensive cloning of sequence tags, overpriced Sanger sequencing (1st generation sequencing) and need of large amount of input RNA have restricted its use (Moody, 2001; Shiraki et al., 2003; Sudhagar et al., 2018; Velculescu et al., 1995).

The development of next-generation sequencing (NGS) has revolutionized the transcriptome profiling. RNA-seq is a modern RNA sequencing technology used to study the entire transcriptome of an organism by converting mRNA into cDNA, and then sequencing the cDNA using NGS. Initially this technique was used to analyse the transcriptome of a eukaryotic organism (a yeast). Although widely used nowadays for the study of expression of human genes, NGS has also been widely applied to study the bacterial transcriptomes (Nagalakshmi et al., 2008; Vliet, 2010). RNA-seq provides

distinct advantages over previous methods. The method is cost effective, provides high coverage and no prior knowledge of the sequence is required but prior knowledge makes alignment much easier. The data obtained is highly reproducible, quantitative, and precise. Recently, several advances have been introduced in the RNA-seq workflow from sample preparation to sequencing platforms and bioinformatics analysis enabling deep profiling of the transcriptome. (Kukurba & Montgomery, 2015; Pinto et al., 2011; Sudhagar et al., 2018). Before describing the RNA-seq workflow, we need to understand Next generation sequencing which is explained in the following section.

5.1.2 Next generation sequencing

NGS is a high throughput DNA sequencing technology widely used for sequencing and assembly of entire genomes (Kim et al., 2012). Roche 454, the first high-throughput sequencing platform, was introduced in 2005. Nowadays, various high throughput sequencing platforms are available like Illumina Hiseq/Miseq, Illumina Novaseq, ABI SOliD, PacBio and Ion Torrent (Hong et al., 2020a; Liu et al., 2012; Marguerat & Bähler, 2010). Each platform has its own pros and cons however, Illumina is currently the key player of NGS. Following are the basic steps of Illumina based next generation sequencing.

1) **Library preparation**

   The sample DNA (genomic DNA or cDNA in case of RNA-seq) is fragmented randomly and specialized adapters are ligated to the 5' and 3' ends of both fragments (Figure 5.1). DNA fragments are now called sequencing templates. A pool of similar sized DNA fragments with known adapters attached to both ends is called a sequencing library. The adapters contain sequences complementary to the lawn of primers present on the flow cell (Figure 5.2). Patterned flow cell is a glass substrate containing billions of nanowells engraved into the surface in a specific array. Each nanowell is seeded with two types of specified primers. Each adapter contains a p5 or p7 sequence for binding to the sequencing templates on the flow cell and Rd1 SP or Rd2 SP primer binding site to start sequencing. In case of multiplexing (sequencing of multiple samples in a single run), index sequences (index 1 and/or index 2) are also added to the adapters for sample identification (Figure 5.1 & Figure 5.2) (Berglund et al., 2011; Metzker, 2010; Quail et al., 2012).

2) Cluster amplification

The double stranded DNA fragments of the sequencing library are denatured and allowed to hybridize to the immobilized primer sequences present in the nanowells of the flow cell through complementary adapter sequences (Figure 5.3-A). DNA polymerase creates complementary strand of the hybridized fragment (this fragment is attached to the flow cell surface) (Figure 5.3-B). The double stranded molecule is denatured, and the original fragment is removed (which was not covalently attached to the flow cell surface). The free end of the DNA fragment interacts with neighbouring primer forming a bridge structure (Figure 5.3-C). Polymerase generates the complementary strand of the fragment forming a double stranded bridge which is then denatured resulting in two ssDNA fragments bound to the surface (Figure 5.3-D & E). This process is called bridge amplification (Goodwin et al., 2016). The latest exclusion amplification methodology allows binding of only a single DNA template to bind a single well. Each nanowell contains hundreds of primers. DNA template after binding to the seeding primer amplifies immediately and rapidly through bridge amplification process to form hundreds of copies of each fragment forming a cluster within a nanowell (Figure 5.3-F & G). The speedy amplification process prevents other templates from binding to the primers of the same well hance preventing polyclonal cluster formation and ensuring a monoclonal cluster formation in each well. The process continues until no unused primers are left in the nanowells. Bridge amplification occurs simultaneously for millions of clusters resulting in clonal amplification of all the fragments of the library (Figure 5.3) (Goodwin et al., 2016; Mardis, 2013; Slatko et al., 2018).

3) Sequence by synthesis (SBS)

After cluster amplification, the reverse strands of the DNA fragments are cleaved and washed off leaving only forward strands of the amplified DNA fragments. The 3' ends are blocked to prevent the DNA fragments from bridging again (Figure 5.3-H & I). To begin the sequencing process, sequencing primers bind to the read 1 sequencing primer (Rd1 SP) of the adapter sequence followed by polymerase binding. In each cycle, a mixture of all the four fluorescently tagged and 3' blocked deoxynucleotides (dNTPs) are added. The single matching base is incorporated to

each complementary strand being elongated and unused nucleotides are washed away (Figure 5.4 - A). The newly added bases are excited by a light source resulting in emission of a characteristic wavelength from each base. The flow cell is imaged by microscopy using laser channels; the colour recorded is used to identify the incorporated base at each cluster. The fluorophore dye and 3' blocking group are removed generating 3'-OH group (Figure 5.4 - B, D). The cycle is repeated until the full length of the growing DNA strand is synthesized. This sequencing technique is called sequence by synthesis (SBS) (Figure 5.4) (Goodwin et al., 2016; Mardis, 2013; Slatko et al., 2018).

For paired end sequencing, the synthesized read 1 fragments are washed away and the 3' ends of the DNA fragments are unblocked allowing them to form bridging again. The reverse strand of each DNA fragment is formed. The forward strand of all the amplified DNA fragments is cleaved and washed off leaving the reverse strand attached to the flow cell. The 3' ends are blocked, and sequencing process is repeated using read 2 sequencing primer (Rd2 SP). Index 1 and index 2 reads are also recorded to identify the libraries (Kukurba & Montgomery, 2015).

Sequencing generates billions of reads (read is the DNA sequence from one fragment). In case of multiplex process, sequence reads for each sample are separated based on the index sequences. Forward and reverse reads are paired to create contiguous sequences.

4) Data Analysis

The contiguous sequences are aligned back to the reference genome for using different bioinformatics tools and software. De novo assembly of the sequences is done when reference genome is not available.

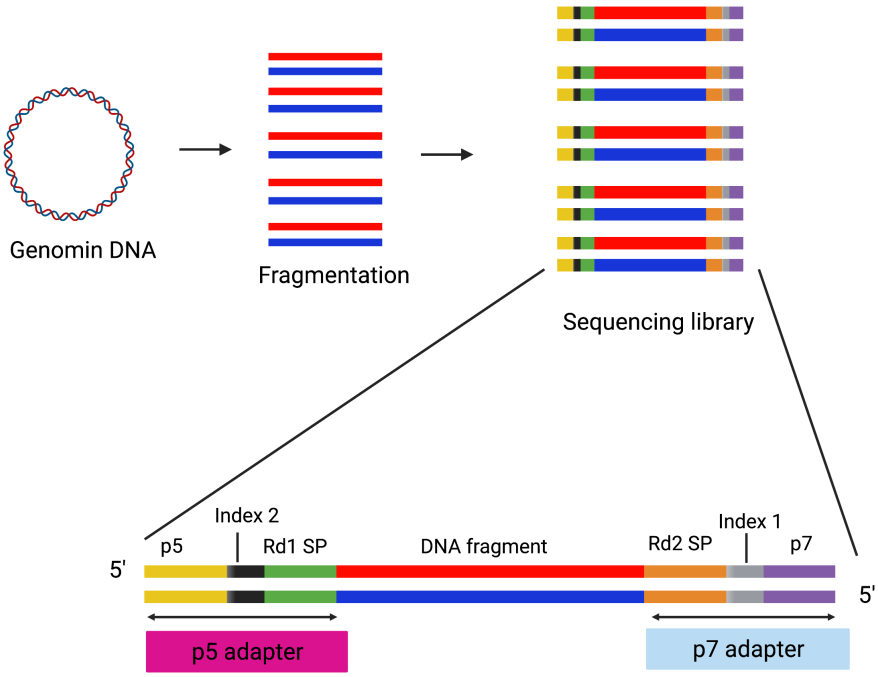


Figure 5.1 Library preparation steps.
The genomic DNA (or cDNA) is fragmented into small pieces to sequence all the fragments in a parallel fashion. The size of the fragments depends on the sequencing method followed. Adapters are ligated to both ends of the DNA fragments. Commonly known adapters are p5 and p7. Adapters have complementary regions for binding to the primer sequences on the flow cell. Rd1 SP and Rd2 SP are the sequencing primers used for sequencing of the DNA fragments. Figure was made using https://biorender.com.

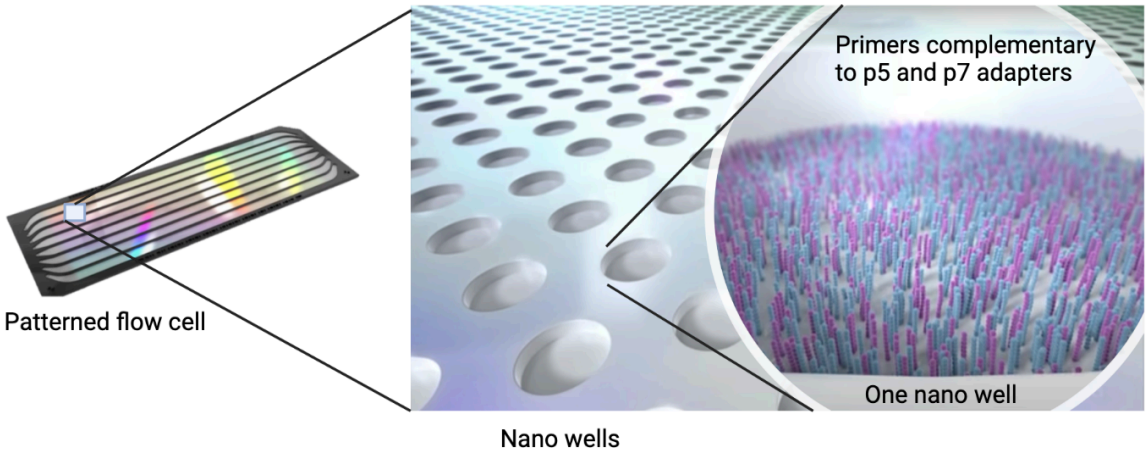


Figure 5.2 Patterned flow cell.
Billions of nanowells are engraved at fixed locations. Each nano well contains hundreds of primers complementary to the adapter sequences (p5 and p7). Image was taken from https://emea.illumina.com.

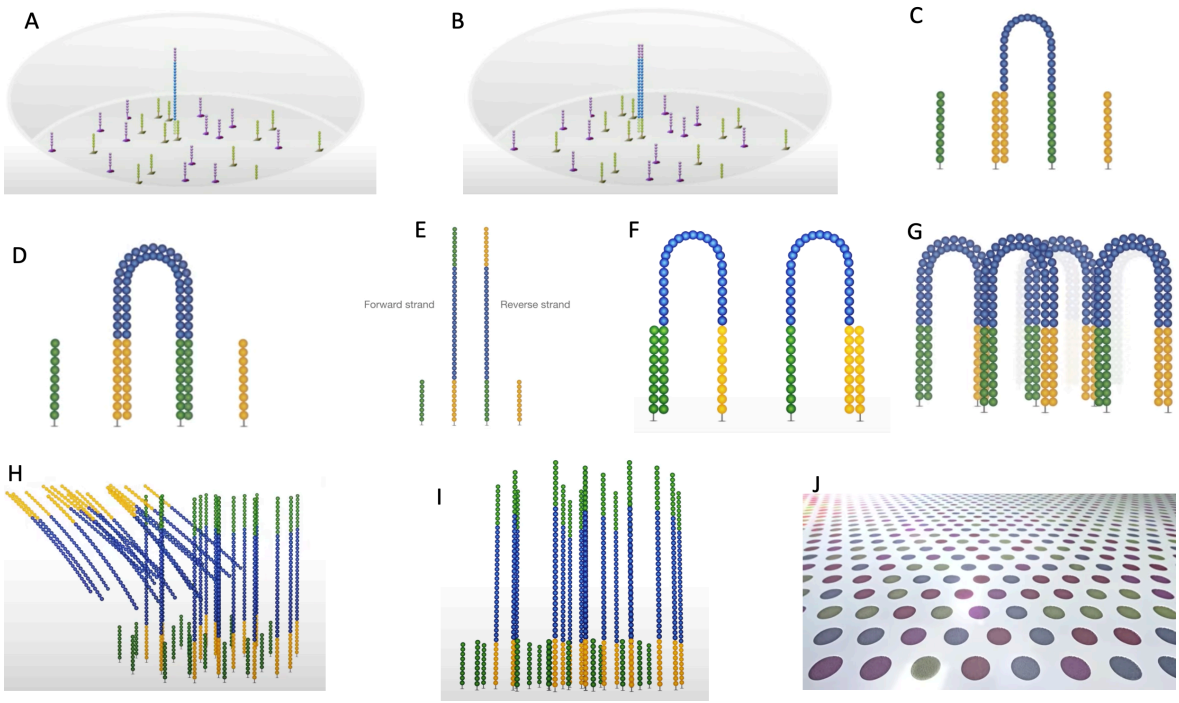


Figure 5.3 Cluster amplification steps.
Single nano well is shown with primers. One DNA fragment binds to the complementary primer through adapter sequence. B) A polymerase enzyme extends the 3' end of the hybridized primer to the full length of the bound DNA fragment. This means flow cell primer is the starting point for building the cluster. The double stranded DNA fragment is then denatured, and original strand is washed away. C) The newly synthesized DNA strand flips over and hybridize to the adjacent complementary primer forming a bridge structure. D, E, F, G) Bridge amplification steps are shown. H) Cleavage and removal of reverse strands. I) 3' ends of the forward strands are blocked. J) Cluster amplification in billions of nano wells occur simultaneously. Images were taken from https://emea.illumina.com and https://www.youtube.com/watch?v=HMyCqWhwB8E.

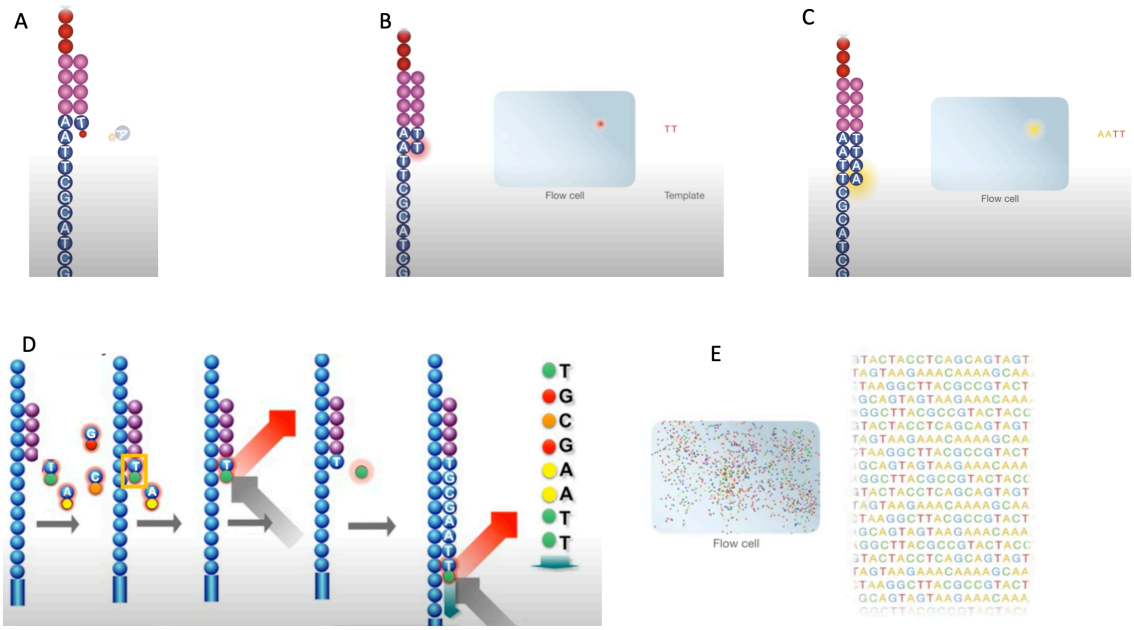


Figure 5.4 Sequencing by synthesis.
A -D) To begin the sequencing, a sequencing primer binds with the sequencing primer of the adapter (Rd1 SP). Four fluorescent dNTPs compete for the growing chain. Only one is incorporated complementary to the sequence of the DNA template. Clusters are excited by a light source after the addition of each nucleotide and a characteristic wavelength is emitted. E) All identical strands are read simultaneously. Clusters are sequenced in a tremendously parallel process. Images were taken from https://www.youtube.com/watch?v=oIJaA6h2bFM&t=861s.

5.1.3 RNA-seq workflow

The key steps involved in transcriptome studies using RNA-seq are RNA extraction, library preparation, sequencing of the library using NGS platform and bioinformatics analysis (Sudhagar et al., 2018). In this section, the steps involved in RNA-seq analysis will be explained briefly.

After careful planning of the experiment, the total RNA is extracted from the bacterial cells. The RNA is quantified using Qubit Fluorometer or NanoDrop Spectrophotometer. Before proceeding further, it is crucial to check the quality of the RNA which is usually done by calculating the RNA integrity number (RIN). RIN values range from 1 to 10, higher the RIN number greater the integrity of the RNA. A RIN number less than 6 is not recommended for RNA-seq analysis because the degraded RNA may result in poor quality and misleading results (Kim et al., 2012; Sudhagar et al., 2018).

Only 5% of the total RNA isolated this way is mRNA; the remainder is mostly rRNA plus small amounts of transfer RNA (tRNA) and other RNAs like small RNAs (sRNAs) (Kim et al., 2012). Ribosomal RNA, as the major contaminant, needs to be mostly removed before RNAseq is done. Prokaryotic mRNA has very short half-life and is highly unstable compared to eukaryotic mRNA. Bacterial mRNA does not have 3' poly A tail which is the characteristic of eukaryotic mRNA where it acts as a stabilizer and signal to protect the mRNA from enzymatic degradation and causes transcription termination (Dreyfus & Régnier, 2002; Kim et al., 2012). To resolve both the issues, mRNA enrichment is a crucial step in prokaryotic RNA-sequencing. Generally used mRNA enrichment methods include rRNA capture/depletion, 3' adenylation, degradation of processed RNA and antibody capture (co-immunoprecipitation) of specific RNAs (Dreyfus & Régnier, 2002). Various kits are available for efficient removal of rRNA such as Ribo-Zero™ technology and NEBNext rRNA Depletion Kit (Kim et al., 2012; Pease & Sooknanan, 2012).

After enrichment, mRNA is fragmented, and reverse transcribed to double stranded cDNA using random primers. Subsequently, both ends of the fragmented cDNA are ligated with specialized adapters in a known orientation. For multiplexing, index sequences are also added for sample identification (Figure 5.1). In some procedures, library enrichment is also performed using limited cycle PCR. After validation of the

sequencing library, the libraries are loaded on the flow cell (Figure 5.5 H-I). Next generation sequencing is performed as described in section 5.1.2 (Figure 5.2-Figure 5.4). Various NGS platforms have their own pros and cons, the choice of the platform generally depends upon the preferred length of the reads, cost, and availability. Short reads (50 -200 bps) provide better coverage and less errors while long reads (>1000 bps) are preferred if reference genome sequence is not available for the organism and de novo assembly is to be performed (Kukurba & Montgomery, 2015; Mantere et al., 2019; Sudhagar et al., 2018). The next generation sequencing platform generates raw sequence data which is converted into FASTQ format files and then used further for bioinformatics analyses as explained in the following section 5.1.4.

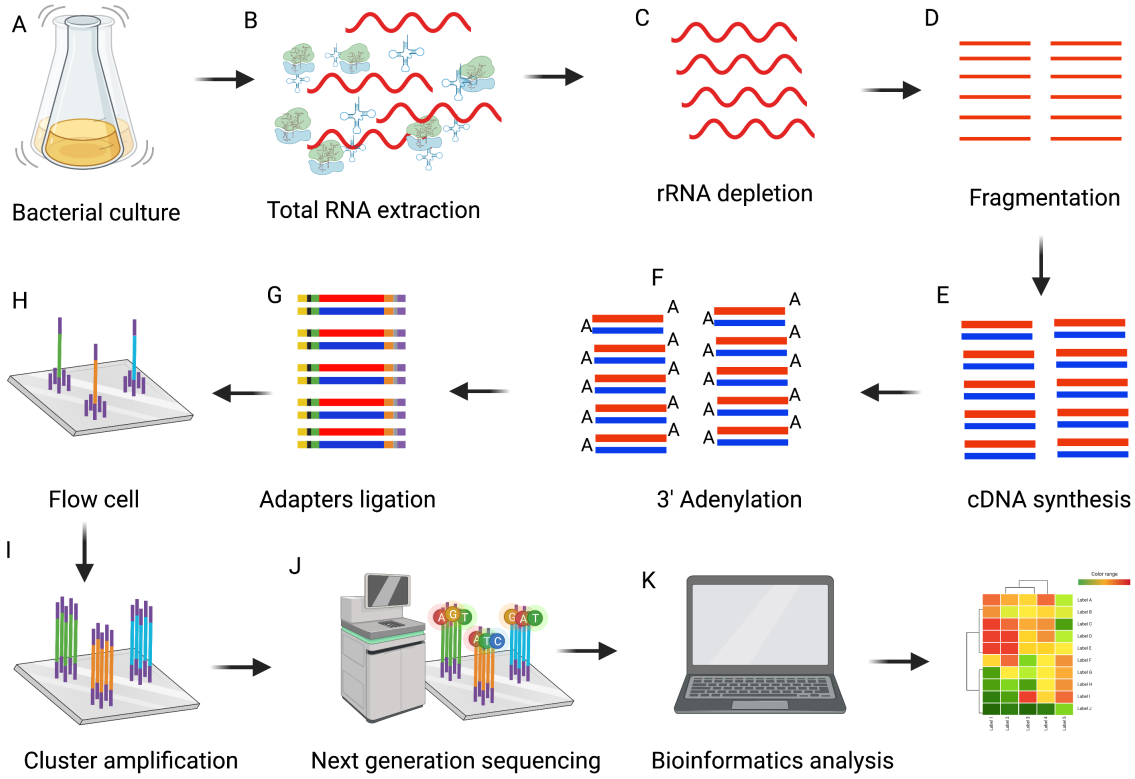


Figure 5.5 Major steps of RNA-seq analysis workflow.
A) Bacterial cultures are grown under specific physiological conditions suitable for the experiment. B) Total RNA is extracted which contains different types of RNAs including mRNA, tRNA and rRNA. C) rRNA depletion is performed. Followed by D-G) fragmentation of RNA, cDNA synthesis, 3' adenylation and adapters ligation. H-I) Library is then loaded on the flow cell where cluster/clonal amplification is performed. J-L) Next generation sequencing is performed, and data is analysed using bioinformatics software. Figure was made using https://biorender.com.

5.1.4 Bioinformatics analysis of RNA-seq data

In the past decade, the bioinformatics analysis tools for RNA-seq have significantly improved. The choice of the bioinformatics analysis pipeline or a computational tool depends on the aims of the experiment and accuracy of the application. The basic steps in any RNA-seq data analysis include quality control of raw reads, alignment of the clean reads to the reference genome, quantification of the aligned reads to quantify the expression levels of genes and discovery of novel transcripts. In case of comparison of multiple samples, normalisation is done followed by differential expression analysis (Kukurba & Montgomery, 2015; Poulsen & Vinther, 2018; Sudhagar et al., 2018).

5.1.4.1 Quality control

The first step in RNA-seq data processing is the quality assessment of raw data which are usually provided in FASTQ format. Quality control (QC) measures if the content of the experimental data is suitable for the subsequent analysis steps. Several QC software applications are available such as NGSQC (Patel & Jain, 2012) and FastQC. FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is commonly used for the quality check of pre-alignment FASTQ files. The FastQC report is generated as an HTML document which reports the basic statistics of the reads such as number of sequences flagged as poor quality, per base sequence quality, sequence duplication levels, overrepresented sequences, and adapter contents (Hong et al., 2020b; Williams et al., 2014). RNA-seQC (DeLuca et al., 2012) is used for the quality control of post-alignment files (Graubert et al., 2021).

5.1.4.2 Data cleaning/Pre-processing

If the data pass the quality control step, the cleaning of the data is performed before alignment; this step is called pre-processing. The data is filtered to trim the adapter sequences, remove contaminants, and eliminate low-quality reads and N bases. Trimmomatic (Bolger et al., 2014), PRINSEQ (Schmieder & Edwards, 2011), and SOAPnuke (Chen et al., 2018) are commonly used tools for data trimming. Trimmomatic is more flexible and efficient, specifically optimized for trimming of Illumina NGS paired-end and single-ended FASTQ data. It identifies and trims adapter

sequences and PCR primers. Trimmomatic is available at http://www.usadellab.org/cms/index.php?page=trimmomatic (Hong et al., 2020b).

Another software, *fastp* (Chen et al., 2018) is an ultra-fast and high-performance tool to perform quality control, trimming, filtering, and base correction on FASTQ files in a single run.

5.1.4.3 Alignment of reads to the reference genome

After quality control and cleaning of the raw data, RNA-seq reads are aligned to the reference genome or to the annotated transcriptome. Commonly used RNA-seq data alignment programmes are Bowtie (Langmead et al., 2009), HISAT2 (D. Kim et al., 2015), TopHat2 (Kim et al., 2013), STAR (Dobin et al., 2013), RUM (Grant et al., 2011), and BWA (Li & Durbin, 2009). Each aligner has its own pros and cons in terms of speed, memory requirement and performance. The choice of the tool depends on the preferences and objectives of the researcher. Different alignment tools generate output files of different formats which complicates the downstream processing (Sudhagar et al., 2018). Therefore, a common alignment format called Sequence Alignment/Map (SAM) format is designed that supports all aligners; SAM is a human readable text file. Binary alignment/Map (BAM) format is the binary representation of SAM format; it retains the same information as SAM (Li et al., 2009). The output files of alignment are usually generated in SAM or BAM format. BAM files are compressed and compact in size; hence, allows fast and more efficient retrieval of data than SAM files. SAM/BAM files contain 11 compulsory fields for each alignment line and a variable optional field. SAMtools is a software package used for positional sorting, indexing, of the SAM/BAM files. SAMtools also hide the low-level details making it possible to use the compressed BAM file directly without the need of uncompressing the entire file (Danecek et al., 2021). Indexing produces a file in *.bham.bai format. This file is used along with the aligned BAM file to visualize the reads in IGV browser if required.

5.1.4.4 Quantification of reads

After alignment, quantification of reads to estimate the levels of gene expression is performed. HTSeq-Count (Anders et al., 2015), LibiNorm (Dyer et al., 2019), and Feature-Count (Liao et al., 2014) are commonly used software for gene-level

quantification of raw counts. For downstream differential gene expression analysis, HTSeq-Count is commonly used for counting the aligned reads against each gene. The script *htseq-count* is designed to consider reads mapping to a single gene only and discard reads overlapping with multiple positions or aligning with more than one gene. Suppose two genes have a stretch of common sequence, one of them is differentially expressed while other one is not, if the reads mapping to both genes are counted for both mapped genes, the extra reads from the differentially expressed gene may mislead about the expression level of the other gene. In case of paired end sequencing data, htseq-count counts fragments (a pair of reads) rather than reads so that the complementary reads originated from single fragment are counted only once (Anders et al., 2015; Conesa et al., 2016; Hong et al., 2020a; Poulsen & Vinther, 2018; Srinivasan et al., 2020).

Several software tools are designed for transcript-level quantification such as Salmon (Patro et al., 2017), Kallisto (Bray et al., 2016), String-Tie (Pertea et al., 2015) and Cufflinks (Trapnell et al., 2010). These are fast and efficient software for assembling the RNA-seq alignment into transcripts. These tools assemble RNA-seq reads into transcripts and estimate the transcript quantification (rather than read/fragment counts) and produce the output file in GTF format. Cufflinks package includes other programs such as Cuffcompare and Cuffdiff. Cuffcompare is used to compare the assembled transcripts (present in GTF output) to the annotated reference genome to find out novel transcripts in the RNA-seq data whereas Cuffdiff is used to identify differentially expressed transcripts (Ghosh & Chan, 2016; Hong et al., 2020a; Pertea et al., 2016).

5.1.4.5 Baseline expression

In RNA-seq analysis, the expression level of each gene is measured by counting the mapped fragments, greater counts represent higher expression. Raw counts cannot be used to compare the expression levels of different genes in one sample because a longer gene will have more fragment counts compared to a shorter gene of equivalent expression levels (Conesa et al., 2016). Therefore, instead of using raw counts, counts are normalized to remove the technical biases such as effects of feature length and library size (sequencing depth) by calculating TPM (transcript per million), RPKM

(reads per kilobase of transcript per million reads mapped) or FPKM (fragments per kilobase of transcript per million reads mapped). RPKM and FPKM are very similar to each other; RPKM is used for single-end RNA-seq where reads are counted while FPKM is used for paired-end RNA-seq where fragments are counted rather than reads. However, TPM is more widely used because of inconsistency of RPKM measurements among samples (Li & Dewey, 2011; Wagner et al., 2012; Zhao et al., 2021).

5.1.4.6 Differential gene expression

RPKM, FPKM and TPM are normalization units suitable to compare the expression levels within a single sample. These measures cannot be used for cross-sample comparisons and to determine the differentially expressed genes under different conditions. RPKM, FPKM and TPM comparison across samples is meaningless because of significant differences in library depth and total RNAs among samples (Zhao et al., 2020). Different conditions (such as presence or absence of a gene or stress) can cause fluctuations in the amount of RNA in cells. Besides, when comparing genes across samples, gene length is assumed to remain the same for the compared genes. If comparing the expression levels of different genes in one sample then the reads must be normalized for gene length as well because different genes have different length and longer genes will have higher reads than the shorter ones.

Trimmed mean of M values (TMM) normalization (Robinson & Oshlack, 2010a), DESeq size factor (Anders & Huber, 2010) and UpperQuartile (Bullard et al., 2010) are the commonly used normalization methods for the differential expression measurements among samples (Smid et al., 2018). These methods don't employ gene length normalization rather account for library size (depth) variation among samples and the fact that some highly differentially expressed genes can affect the count distribution negatively impacting the normalization procedure (Conesa et al., 2016; M. D. Robinson & Oshlack, 2010b; Y. Zhao et al., 2021).

Different tools are used to perform the differential gene expression (DGE) analysis such as DESeq2 (Trapnell et al., 2012) and edgeR (Robinson et al., 2010). These tools perform statistical tests on the read/fragment counts or transcript counts derived from the computational analysis described in section 5.1.4.4 to determine the statistically significantly expressed genes. The analysis also provides information about the

expression level of each gene and the pairwise magnitude of difference among different samples cultured under different conditions (McDermaid et al., 2019; Seyednasrollah et al., 2015).

In DESeq2 analysis, DESeq size factor is used to normalize the counts by calculating the geometric mean for each gene present in all samples. It then divides the counts for a gene in all the samples by this mean. DESeq2 uses negative binomial model and Wald test to identify the differentially expressed genes between two samples (Love et al., 2014).

The final output file of the DESeq2 analysis contains p-value (calculated using Wald test), p-value adjusted (padj) and log2 fold change for each gene in the two samples compared. In any statistical analysis, p-value is a measure of the likelihood that the changes did occur by chance. So, smaller the p-value the more likely that they did not. For two samples, (control vs test) null hypothesis states that there is no differential expression for each gene across the samples. In DESeq2 analysis p-values are calculated using the Wald test. The null hypothesis for the Wald test states that there is no differential expression of a gene across the two samples (in test vs control). When p-value is <0.05, the null hypothesis is usually rejected because there is only 5% chance of null hypothesis being true. However, for multiple testing, we need to correct the p-value because the probability of a false positive increases. DESeq2 adjusts the p-value using Benjamini and Hochberg method (Benjamini & Hochberg, 1995) hence we use padj values in analysis rather than p-values (Jafari & Ansari-Pour, 2019). A summary of RNA-seq bioinformatics data analysis workflow is given in Figure 5.6.

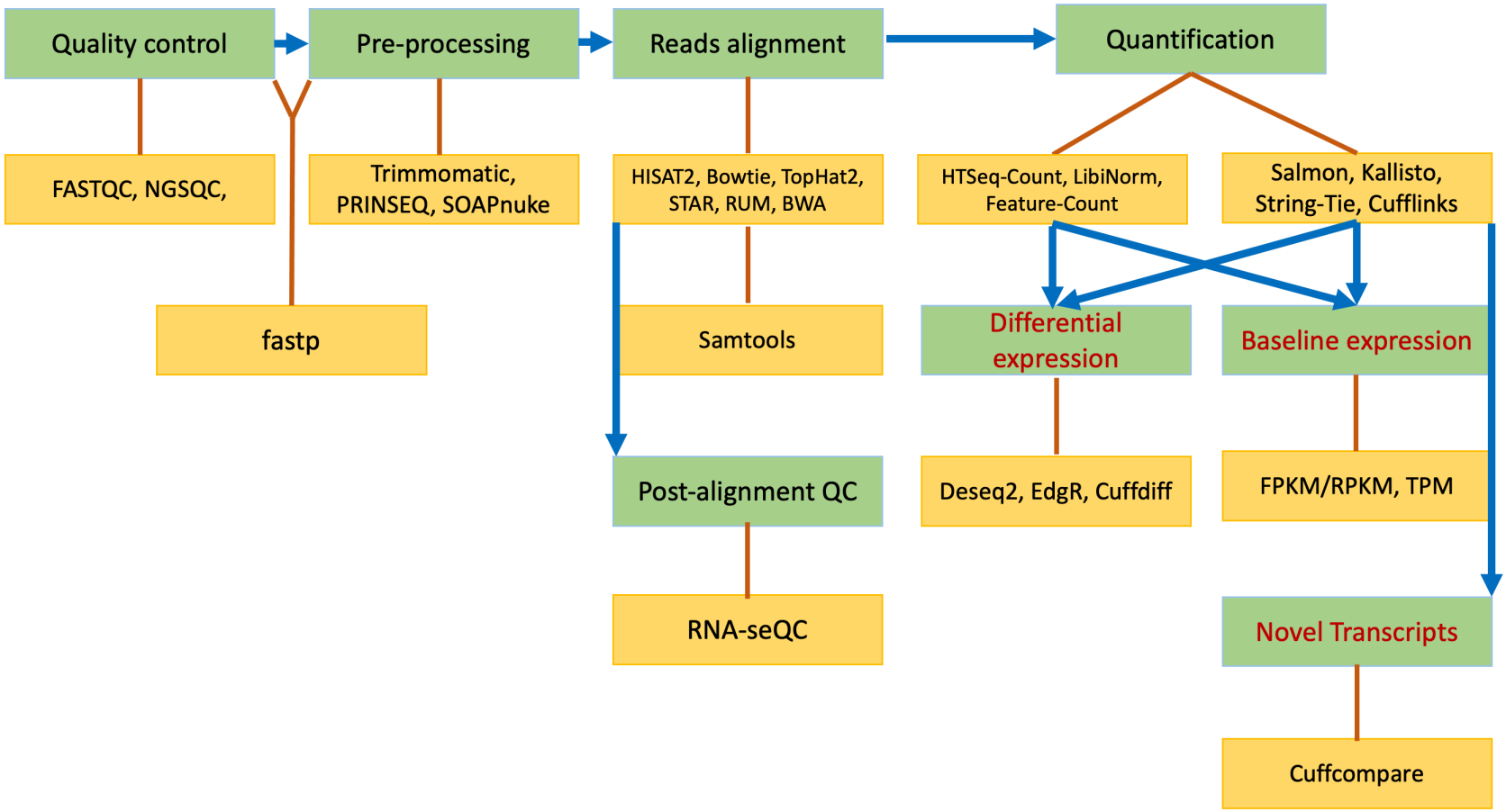


Figure 5.6 Step wise workflow of RNA-seq bioinformatics data analysis.

5.2 Experimental plan

The aim of RNA-seq analysis was to study the expression profile of all the genes in the W.T, *cpn60.1* knockout and complemented strains of *M. marinum* 1218R. It is not known whether Cpn60.1 performs chaperonin function or what is the effect of deletion of *cpn60.1* on the overall transcriptome profiling of the cell. Our working hypothesis was that if Cpn60.1 has any chaperone activity then we expect that the deletion of *cpn60.1* could lead to increased expression of other chaperone genes, including *cpn60.2*, to compensate for its loss. In other words, if the absence of Cpn60.1 cause an increase in the expression levels of other chaperones, then this might be due the missing chaperone role or stress response function of Cpn60.1. To test this hypothesis, *cpn60.1* knockout strain *M. marinum* 1218R *Δcpn60.1* was included in the study.

In addition, the complemented strain was included to cross check the results. Triplicate biological samples of each strain (WT, *Δcpn60.1* (KO) and complemented (COM)) were used for the analysis. The DESeq 2 analysis comparison was done between KO (test/treated) and WT (control), complemented stain (test) and KO (control), and complemented strain (test) and WT (WT). If a gene is observed to be significantly upregulated in the *Δcpn60.1* strain (compared to WT) and the reason of increase of the expression level of that gene is the direct or indirect effect of absence of Cpn60.1, then putting the *cpn60.1* back into the knockout cells (plasmid expressed copy) would cause a significant decrease in the expression level of that gene in the complemented strain when compared to the KO. And there will be no significant up or down regulation of that gene when the differential expression analysis is done between complemented and WT strains. Same is true for vice versa i.e., if significant downregulation of a gene is observed in KOvWT DESeq 2 analysis and the same gene would show downregulation in the COMvKO comparison. Therefore, the genes showing significant upregulation in KOvWT but significant downregulation in COMvKO and vice versa were sorted and categorized according to their functions.

Both the knockout and complemented strains were constructed by Dr C. M. Santosh Kumar in our lab. The *M. marinum Δcpn60.1* was constructed by the swapping of *cpn60.1* with a hygromycin resistance cassette and *sacB* gene (as shown in Figure 5.8), whereas the complemented strain was constructed by introducing pSK101-*Mmcpn10*-

cpn60.1 recombinant plasmid (complementation plasmid, Figure 5.7-A) into the *M. marinum ⊿cpn60.1* strain using electroporation. The plasmid pSK101-*Mmcpn10-cpn60.1* contained a kanamycin resistance cassette. To grow all the three strains under similar physiological condition, pSK101 plasmid without carrying *cpn* genes (vector only, Figure 5.7-B) was introduced in the W.T and KO strains. A symbolic representation of all the three stains is given in Figure 5.9. *M. marinum* 1218R contains a naturally occurring plasmid pMMRN.

RNA samples were sent to a next generation sequencing company, Genewiz ngs.europe@genewiz.com. For next generation sequencing, Illumina Novaseq 6000 instrument was used. For better coverage, short reads of 2 x150 bps paired ends sequencing with 10 M reads per sample was chosen to perform.

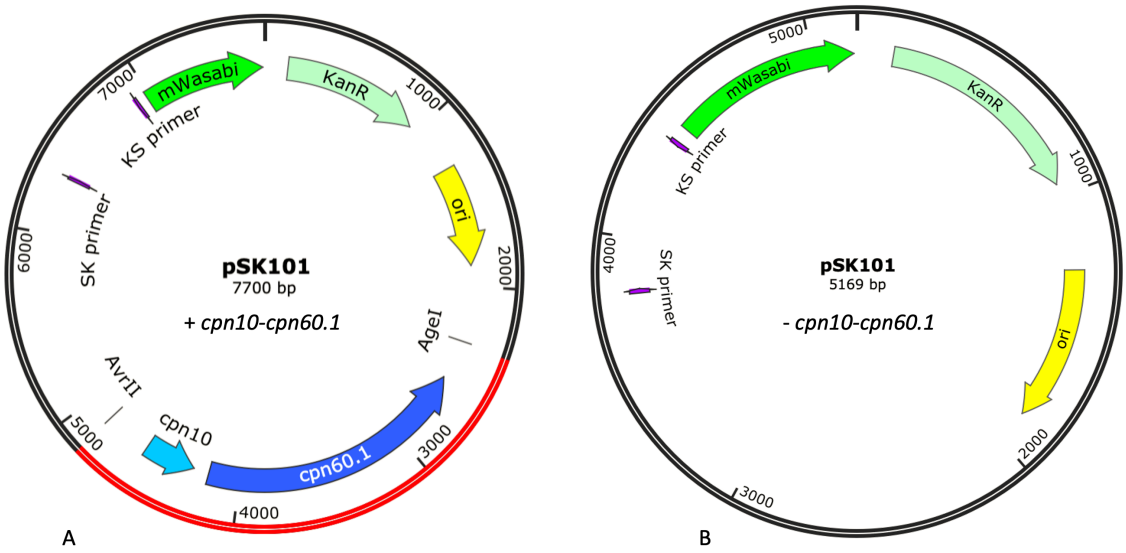


Figure 5.7 pSK101 plasmids (A) with and (B) without *cpn10-cpn60.1*.
A) Complementation plasmid B) Vector only. Both plasmids have a Kan^R cassette.

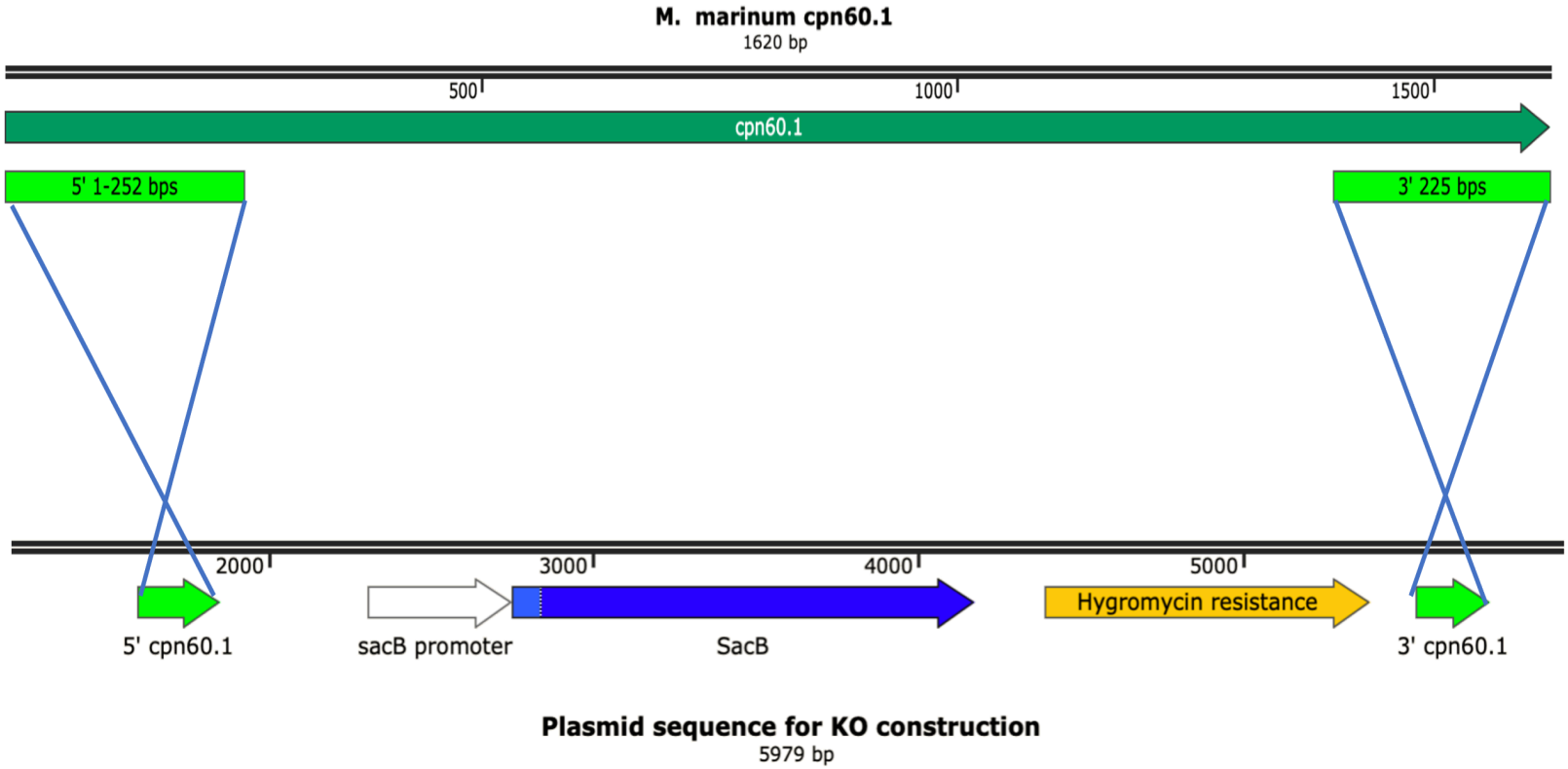


Figure 5.8 Construction of knockout mutant strain *M. marinum*.
Swapping of chromosomal *cpn60.1* gene of *M. marinum* 1218R with hygromycin cassette to construct the knockout mutant strain *M. marinum Δcpn60.1*.

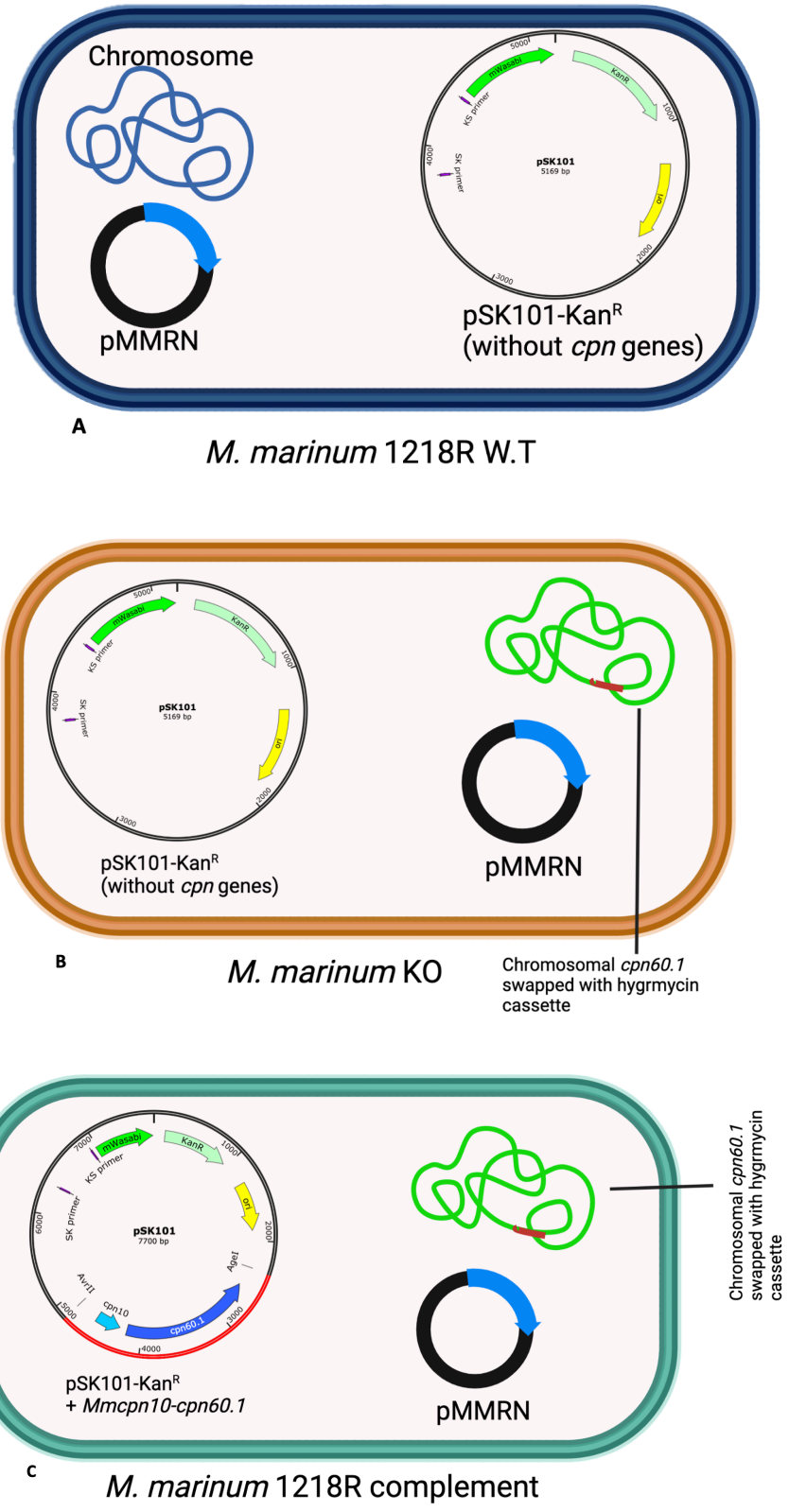


Figure 5.9 Symbolic representation of *M. marinum* W.T (A), KO (B), and complemented strain (C).
Figures were made using https://biorender.com.

5.3 Results

The 2X secondary growth cultures of triplicate biological replicates of all the three strains of *M. mairnum* 1218R (W.T, KO, and Complemented strain) were grown in 7H9 media at 30°C 180 rpm till OD_{600} reached between 0.8 – 1.0. Total RNA from all the samples was extracted using Qiagen RNeasy kit as explained in chapter 2 (section 2.4.11), the elution of RNA was done in nuclease free water.

5.3.1 RNA quantification and integrity check

The RNA samples were quantified using UV spectrophotometer nanodrop settings at RNA-40, λ_{230} nm. RNA integrity number (RIN) was checked using Agilent 2200 TapeStation System using the standard protocol as explained in Chapter 2 (section 2.4.11.1). The RNA samples having integrity number (RIN) less than 6.0 are not suitable for RNA sequencing and higher the RIN more suitable the sample is for RNA-seq analysis, therefore the samples having highest values of RIN numbers were selected (Table 5.1, Figure 5.10). 50 ml (5-10mg) of each sample (9 samples in total) was sent to Genewiz NGS sequencing company for RNA sequencing.

No.	Samples 1X	A_{230} (ng /ml)	Dilution for RIN number calculation	RIN	No.	Samples 2X	A_{230} (ng /ml)	Dilution for RIN number calculation	RIN
1	WT_1	1032.8	3.5	6.1	10	WT_1	1717.5	5.1	5.2
2	WT_2	1217.4	3.7	6.9	11	WT_2	1180.7	2.7	5.4
3	WT_3	740.1	4.3	7.1	12	WT_3	1792.5	4.8	6.2
4	KO_1	1778.1	5.8	5.4	13	KO_1	1810.0	5.7	7.3
5	KO_2	598.6	4.2	5.4	14	KO_2	159.1	1.7	7.7
6	KO_3	1976.0	7.3	5.9	15	KO_3	2468.2	7.1	7.5
7	COM_1	970.7	2.7	7.0	16	COM_1	2188.8	6.5	6.9
8	COM_2	1399.3	5.0	7.7	17	COM_2	2674.6	9.0	7.5
9	COM_3	1878.3	6.7	7.5	18	COM_3	2277.5	7.6	7.4

Table 5.1 Table showing the absorbance and RIN number.

The absorbance and RIN number of 2X triplicate samples of WT, KO, and COM strains of *M. marinum*. Green highlighted samples of WT, orange highlighted samples of KO, and yellow highlighted samples of COM were chosen for RNA-seq analysis.

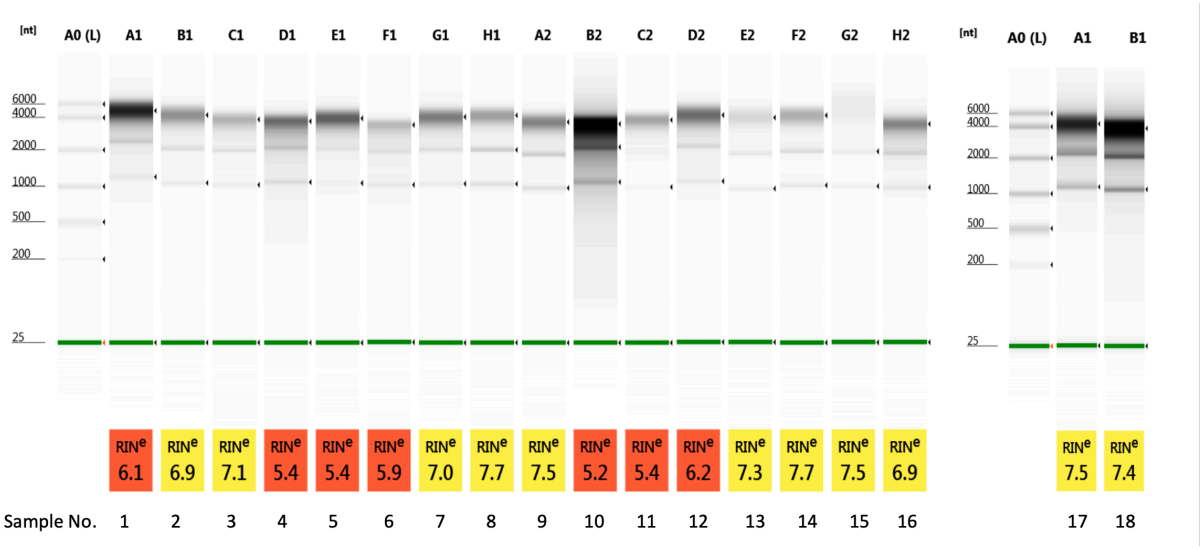


Figure 5.10 TapeStation analysis report showing the integrity numbers of RNA samples.

5.3.2 RNA sequencing – performed by the Genewiz company

The company performed the DNase treatment of all the samples to remove any contaminating DNA. NEBNext rRNA Depletion Kit (Bacteria) was used to perform rRNA depletion. RNA sequencing library was prepared using NEBNext Ultra II RNA Library Prep Kit for Illumina using manufacturer's protocol (NEB, Ipswich, MA, USA). Briefly, the enriched RNA samples were fragmented, and double stranded cDNA was synthesized (Figure 5.11 A-C). cDNA fragments were end repaired, 3' ends adenylation was performed and adapters were ligated followed by index addition (Figure 5.11 D & E). Library was enriched with limited cycle PCR. The libraries were validated using NGS Kit on the Agilent 5300 Fragment Analyzer (Agilent Technologies, Palo Alto, CA, USA), and quantified by using Qubit 4.0 Fluorometer (Invitrogen, Carlsbad, CA).

The sequencing libraries were multiplexed and loaded on the flow cell on the Illumina NovaSeq 6000 instrument according to manufacturer's instructions. The samples were sequenced using a 2x150 Pair-End (PE) configuration v1.5. Image analysis and base calling were conducted by the NovaSeq Control Software v1.7 on the NovaSeq instrument. Raw sequence data (.bcl files) generated from Illumina NovaSeq was converted into FASTQ files and de-multiplexed using Illumina bcl2fastq program version 2.20. One mismatch was allowed for index sequence identification. The quality of the raw data was investigated, and sequence reads were trimmed to remove possible adapter sequences and nucleotides with poor quality using Trimmomatic v.0.36. The

raw data was delivered as two paired end sequence files R1 and R2 for each sample in the FASTQ file format.

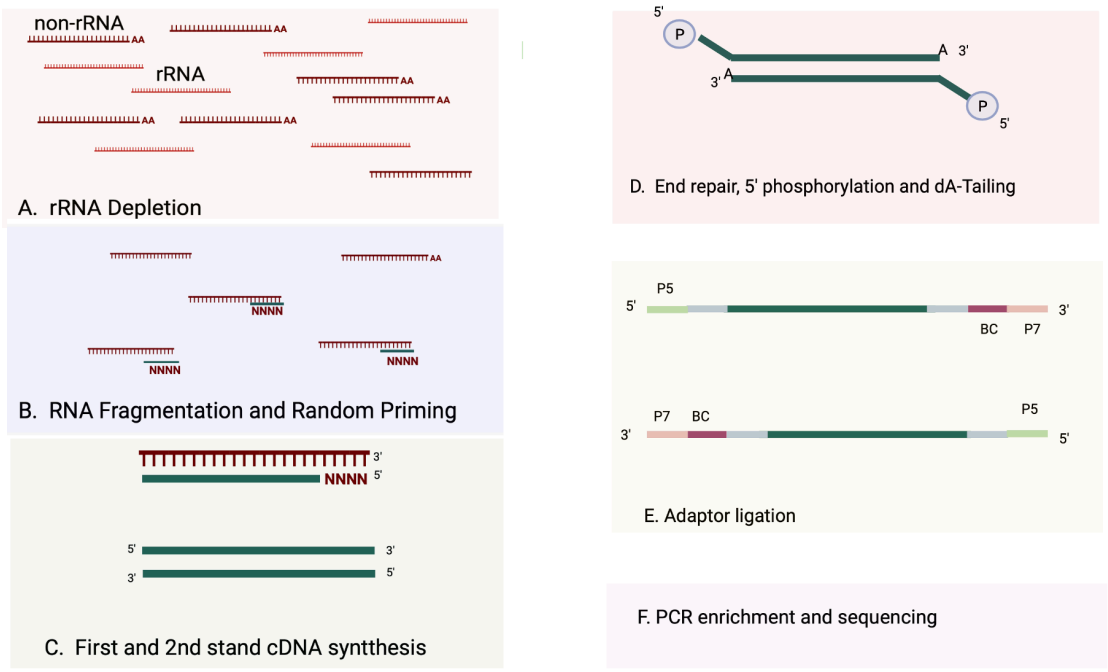


Figure 5.11 Different steps involved in RNA seq library preparation.
Image produced using https://biorender.com.

5.3.3 Bioinformatics analysis of RNA-seq data

The bioinformatics analysis on the raw data provided by Genewiz was performed by me as described below. The data was provided in two fastq zip files (R1 and R2) for each sample.

5.3.3.1 Quality control using FastQC

A very popular software FastQC was used for quality control check of the raw data (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). All the 18 files of raw data of 9 samples were individually analysed.

The sequence length of all the reads among all samples was between 8-150 bases. The per base sequence quality plots of four files RNA-seq raw data are shown in Figure 5.12. Overall, all the files showed similar quality of reads. The x-axis of the graph shows the position of base along all the reads in the file. In the start of the graph, each position is shown separately from 1 to 9 then the positions are binned together in 5 bps. So, the

plot after 9 bases shows the collective statistics for 5 bps. On y-axis quality scores are shown. The y-axis is divided in three coloured backgrounds depending on the quality of the scores. Green area represents the very good quality base calls. The middle orange part represents reasonable quality calls and red area shows the poor-quality calls. The blue line is showing the mean value of the quality score at each position or window of the bps. The distribution of average read quality score of all the bases was found to be very high for all the 18 files of 9 samples. No file showed presence of any adapter sequences in the reads. Figure 5.13 shows the adapter content plot of WT_1, R1 file. The plots of all other files were similar to this plot.

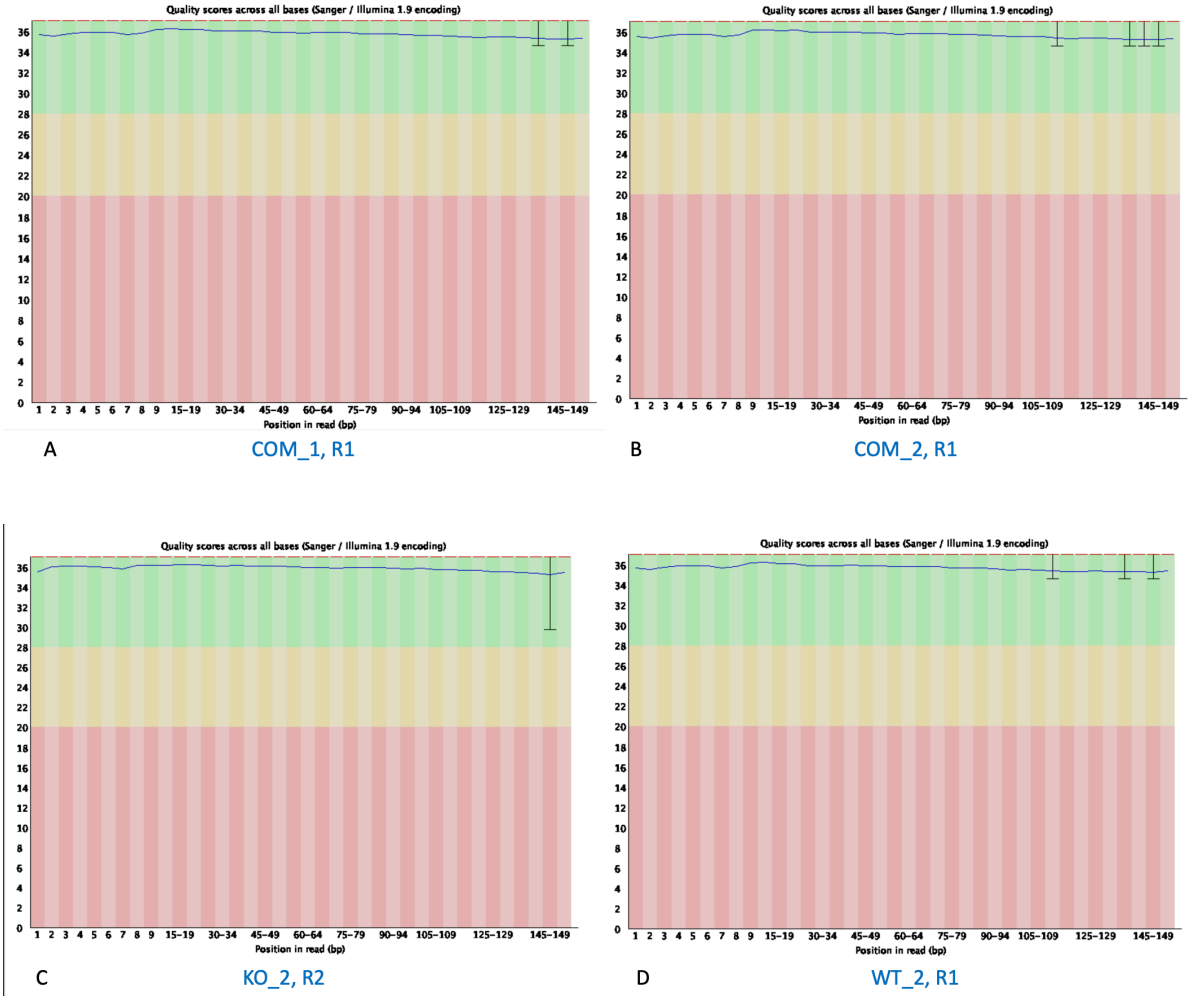


Figure 5.12 Per base sequence quality of the four files of RNA seq raw data.

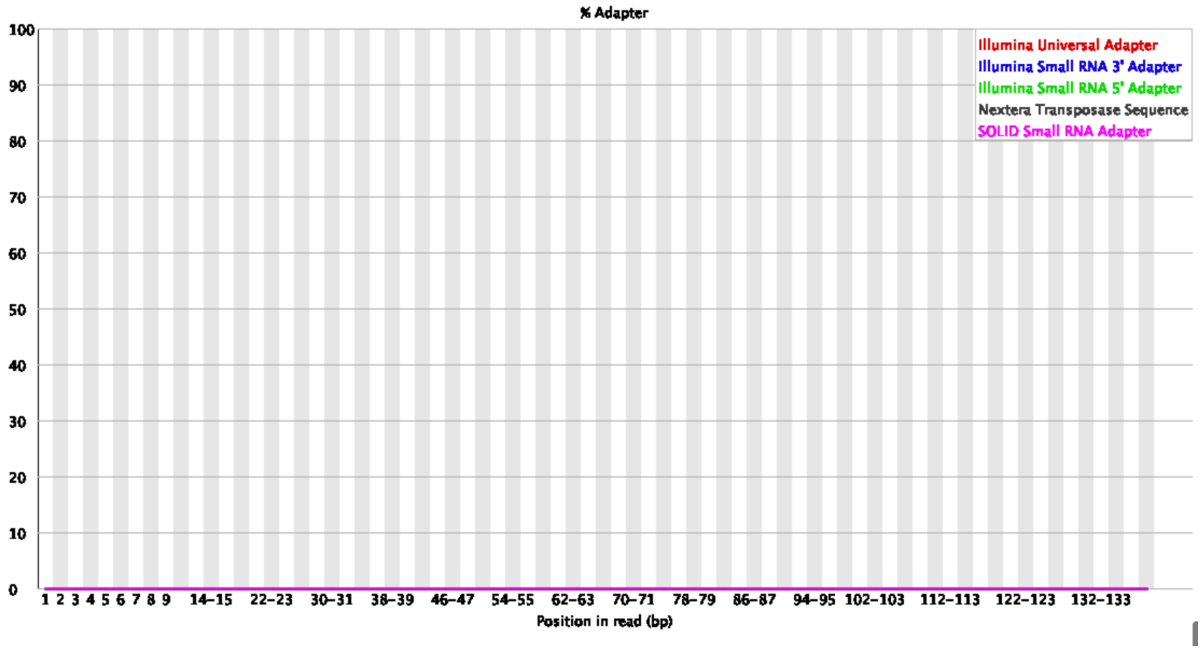


Figure 5.13 Adapter content plot of WT_1, R1 file.

5.3.3.2 Alignment/Mapping of the sequencing reads to the reference sequences

Alignment of the sequencing reads was done against the FASTA sequence files of the reference sequences. The symbolic diagram of the three strains of *M. marinum* 1218R used for RNA-seq analysis is shown in Figure 5.9. Two genome reference sequence files were used. For the alignment of the reads of WT samples, the complete chromosome reference sequence file was downloaded from NCBI (Accession no. NZ_CP025779.1).

Because the middle part of the *cpn60.1* gene in the chromosome of both the KO and COM strains was swapped with *sacB* and hygromycin resistance cassette (see Figure 5.8), therefore the FASTA sequence of the WT genome downloaded from NCBI was modified at the *cpn60.1* gene position. The sequence of the swapped region of *cpn60.1* was removed and the FASTA sequences of *sacB* and *hygromycin* genes were added in that place. This modified FASTA file was used for the alignment of sequence reads KO and COM samples.

All the three strains contain a native plasmid pMMRN, therefore the sequence reads of all the samples were also aligned against FASTA sequence of pMMRN DNA. The sequence of the plasmid was downloaded from NCBI (Accession no. AP018497.1). All

reads files of all the samples were also aligned against the FASTA sequence file of pSK101 plasmid sequence for the complemented strains.

Alignment was done using the HISAT2 mapping program version 2.2.1. The number of reads and overall alignment rate for all the samples is given in Table 5.2. Except for the WT_1 sample, all the samples showed overall alignment rate of more than 90% against the reference genome.

5.3.3.3 SAM to BAM conversion, sorting, and indexing

The alignments produced by HISAT2 are in random order and the out file was produced in SAM format. It is advantageous to sort the SAM files so that the alignments occur with respect to their position in the reference genome as describe in section 5.1.4.3. Therefore, the positional sorting of SAM files and conversion from SAM to BAM format was done using SAMtools. Sorted BAM files can be used to view in the genome viewer browsers against the reference sequence (Figure 5.14). Indexing of the sorted BAM files was also performed using SAMtools index.

No.	Sample	No of reads	Overall Alignment rate against genome	Overall Alignment rate against pMMRN	Overall Alignment rate against pSK101
1	WT_1	14092653	78.97%	0.12%	0.17%
2	WT_2	8908925	95.47%	0.13%	0.31%
3	WT_3	13528499	94.14%	0.11%	0.19%
4	KO_1	11820367	96.48%	0.18%	0.79%
5	KO_2	20570921	96.59%	0.14%	0.53%
6	KO_3	12258742	96.08%	0.21%	1.11%
7	COM_1	22455848	93.49%	0.09%	1.13%
8	COM_2	9043901	95.37%	0.10%	1.25%
9	COM_3	37169114	94.56%	0.07%	0.79%

Table 5.2 Number of reads and overall alignment rate.
Table showing the number of reads of all the samples and overall alignment rate against the reference sequences i.e., genome, pMMRN, and pSK101.

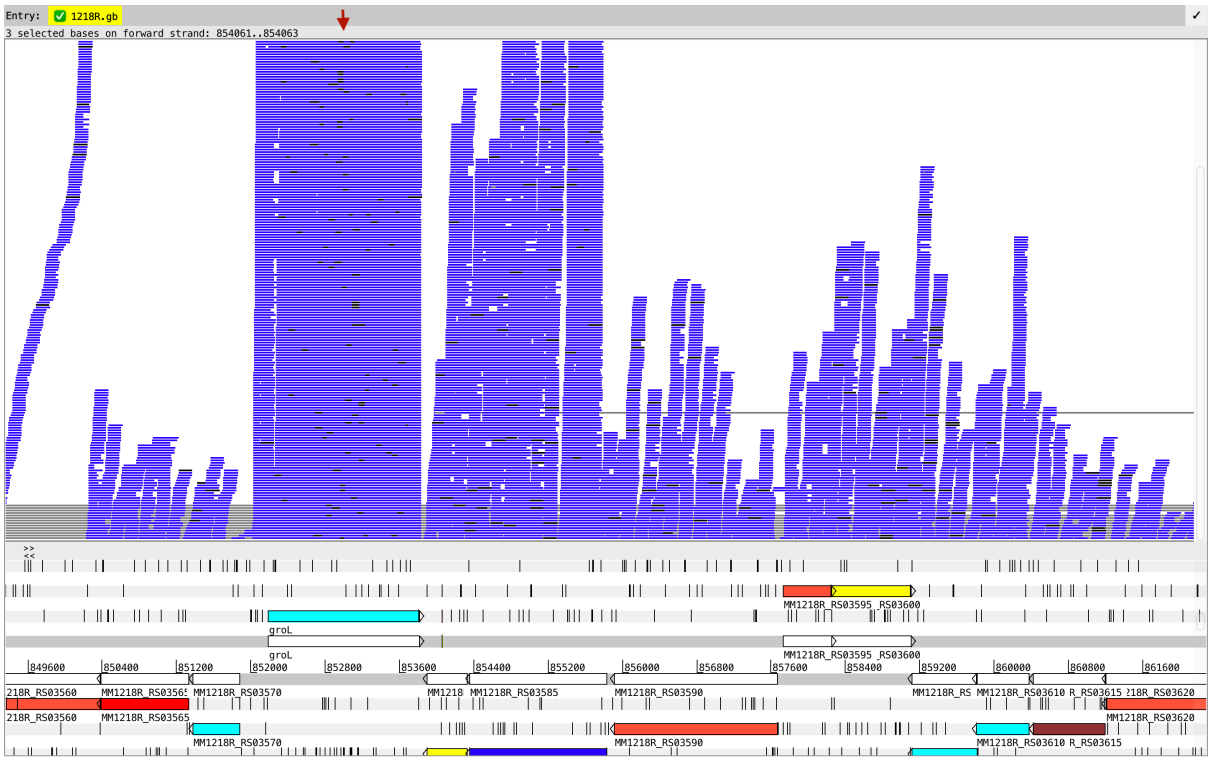


Figure 5.14 BAM file view in Artemis genome browser.
The sorted BAM file of WT_3 (sample no. 3) opened in Artemis genome browser. The area of the reads aligning to the *cpn60.2* (*groL*) is shown.

5.3.3.4 Quantification of reads

After sorting and indexing of the mapped reads, quantification of the reads was done using HTSeq-count program. The input files of HTSeq-count are aligned sequencing reads (aligned.bham file of each sample) file and a file containing information of the genomic features such as the gff3 file of the reference sequence.

For W.T samples, gff3 annotation file of *M. marinum* 1218R (Accession no. NZ_CP025779.1) was downloaded from NCBI. For the samples of KO and COM strains, the gff3 file (of WT genome) was modified. The annotation for *cpn60.1* was replaced with 4 annotations. The 5' *cpn60.1* flanking region was annotated as MMgroELF_001, the *sacB* gene as MMSacB_002, the hygromycin resistance cassette as MMHygromycin_003 and the 3' flanking region as MMgroELR_004. pMMRN plasmid annotation file was downloaded from NCBI (Accession no. AP018497.1) and the annotation file for pSK101 plasmid was created using Prokka.

All the aligned bam files were used for quantification of reads against genome annotation files and plasmids annotation files. The output of the Htseq-count was a text file containing first row of gene ID or locus tag of the coding sequence and a second row of number of reads counted against each coding sequence. The counts from all the samples for genome and plasmids were combined in one csv file. The counts for the first 25 genes are shown in Figure 5.15. This is the raw counts file which cannot be used to draw conclusions directly because the counts are required to get normalized before performing any statistical analysis.

	A	B	C	D	E	F	G	H	I	J
1		WT_1	WT_2	WT_3	KO_1	KO_2	KO_3	COM_1	COM_2	COM_3
2	MM1218R_RS00005	368	453	491	600	633	715	640	316	954
3	MM1218R_RS00010	1250	848	1044	1713	2443	2056	1418	720	2072
4	MM1218R_RS00015	620	360	449	814	1299	1175	612	234	900
5	MM1218R_RS00020	420	256	299	589	901	790	400	192	626
6	MM1218R_RS00025	2348	1831	2099	3808	4601	5574	2994	1204	3279
7	MM1218R_RS00030	2429	1878	2050	4537	5643	5981	2438	1023	3222
8	MM1218R_RS00035	967	701	948	1961	2560	2615	1325	537	1970
9	MM1218R_RS00040	49	59	46	36	65	47	50	31	92
10	MM1218R_RS00045	95	95	118	147	321	218	147	62	170
11	MM1218R_RS00050	185	84	144	252	387	410	247	114	341
12	MM1218R_RS00055	117	107	140	152	226	230	169	86	263
13	MM1218R_RS00060	368	213	412	403	559	503	531	262	864
14	MM1218R_RS00065	874	729	897	1262	1294	1546	1376	525	1774
15	MM1218R_RS00070	141	136	161	175	199	275	224	96	281
16	MM1218R_RS00075	141	129	194	273	279	275	203	109	267
17	MM1218R_RS00080	88	96	99	124	120	137	150	44	173
18	MM1218R_RS00085	360	207	267	503	611	648	376	155	545
19	MM1218R_RS00090	1060	1018	1106	1496	1672	1919	1533	641	2159
20	MM1218R_RS00095	677	556	584	911	1075	1230	934	374	1180
21	MM1218R_RS00100	942	748	739	1306	1493	1681	1139	498	1634

Figure 5.15 The raw read counts generated by HTSeq-count. Only the first 21 genes are shown.

5.3.3.5 Differential gene expression analysis

The raw read counts were normalised using DESeq2 R package. DESeq2 normalized the raw counts against library size. The normalised read counts were then used for further statistical analysis of the samples which is discussed in the next section 5.3.4. DESeq2 was also used for the differential gene expression analysis. The analysis was done between KO (test) and WT (control) (KOvWT), COM (test) and KO (control) (COMvKO), and COM (test) and WT (control) (COMvWT). A sample of the files generated by the DESeq2 analysis is shown in Figure 5.16. The file contains the log2FoldChange and padj values of each gene.

	A	B	C	D	E	F	G
1		baseMean	log2FoldChai	lfcSE	stat	pvalue	padj
2	MM1218R_RS00005	557.805245	-0.3747363	0.20933968	-1.790087	0.073439883	0.127892157
3	MM1218R_RS00010	1395.06537	0.06794086	0.11304306	0.6010175	0.547828325	0.651213235
4	MM1218R_RS00015	635.964217	0.28713165	0.13264578	2.1646497	0.030414521	0.060678705
5	MM1218R_RS00020	444.672387	0.31017373	0.13281148	2.3354436	0.01952026	0.04211464
6	MM1218R_RS00025	2812.49562	0.22989918	0.14854114	1.5477139	0.121691204	0.195204447
7	MM1218R_RS00030	2883.79092	0.42239446	0.11647801	3.626388	0.000287413	0.001032209
8	MM1218R_RS00035	1329.39487	0.53196651	0.10147903	5.2421324	1.59E-07	1.16E-06
9	MM1218R_RS00040	53.6077933	-1.0240394	0.28343789	-3.612923	0.000302764	0.001078482
10	MM1218R_RS00045	138.639835	0.19002729	0.22567974	0.842022	0.399775664	0.51213089
11	MM1218R_RS00050	214.828249	0.44463916	0.17167579	2.5899933	0.009597779	0.023026769
12	MM1218R_RS00055	156.515453	-0.1955503	0.1745159	-1.12053	0.262488038	0.364069451
13	MM1218R_RS00060	433.806876	-0.335785	0.15429857	-2.176203	0.02954011	0.059091482
14	MM1218R_RS00065	1078.57086	-0.1982927	0.15091165	-1.313965	0.188857855	0.280321687
15	MM1218R_RS00070	180.635966	-0.366618	0.19705455	-1.86049	0.062816258	0.111993512
16	MM1218R_RS00075	198.232055	-0.0714963	0.20776075	-0.344128	0.730750151	0.802236117
17	MM1218R_RS00080	110.205923	-0.4972903	0.2341363	-2.123935	0.033675571	0.066005126
18	MM1218R_RS00085	369.073491	0.18156968	0.13357564	1.3593023	0.174050812	0.261824716
19	MM1218R_RS00090	1328.97867	-0.2578072	0.1559305	-1.653347	0.098260261	0.163465945
20	MM1218R_RS00095	783.608054	-0.1024928	0.1416854	-0.723383	0.469444711	0.579570917
21	MM1218R_RS00100	1054.0111	-0.0391981	0.14390272	-0.272393	0.785319562	0.84403354

Figure 5.16 Output file of DESeq2 analysis between KOvWT. Only the first 21 genes are shown.

The default function of DESeq2 was used to create the MA (ratio intensity) plots of each comparison. The MA plots displaying the mean values of normalized counts of all the genes on x-axis against the log2FoldChange on y-axis are shown in Figure 5.17. In all the three comparisons the genes showing a significant change in expression according to the Wald significant test and adjusted p-value < 0.1 are highlighted in blue. The genes having the log2FoldChange value greater than 2 or less than -2 are plotted as triangles. As expected, many genes showed significant up and down regulation in KOvWT comparison (Figure 5.17-A). Similarly, in COMvKO comparison nearly equal number of genes showed significant expression (Figure 5.17-B). These results indicate that the deletion of *cpn60.1* has significant direct or indirect effect on the regulation of many genes in the cell. The effect is reversed by putting the *cpn60.1* gene back (expressed by the plasmid copy) into the knockout strain. Whereas, in COMvWT comparison, only a few genes were observed to show significant expression (Figure 5.17-C) which denote that both the strains behave nearly similar. This again indicates that if any change was caused due to the deletion of chromosomal *cpn60.1* gene then expression of *cpn60.1* from the plasmid reversed the effect of deletion of *cpn60.1* so that the expression levels of nearly all the genes (except a few shown in blue) returned to the WT level.

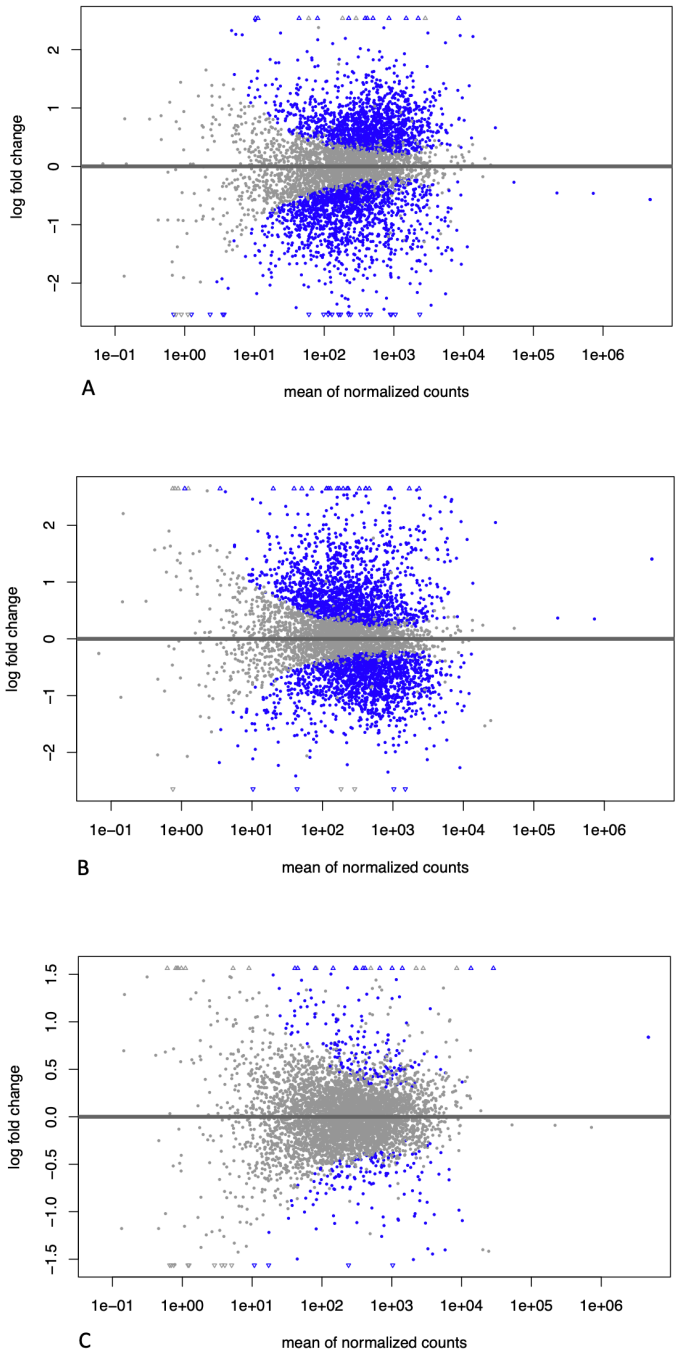


Figure 5.17 MA plots of all the three comparisons; KOvWT (A), COMvKO (B), and COMvWT (C). Blue points show the significantly differentially expressed genes. Created in R studio using DESeq2 package.

5.3.4 Analysis of DESeq2 results, interpretation, and visualization of data

To observe the relationship between two sets of samples, the normalized counts files were used to draw the scatter plots. To visualize the gene expression data, clustered heatmaps were also generated.

5.3.4.1 Scatter plots of normalized counts

The normalized counts of the individual replicates were plotted against each other to look for any cases where there were large differences between replicates of a sample. Ideally, the values of the replicates should corelate with each other. Figure 5.18 to Figure 5.22 show scatter plots of all the possible comparisons of between triplicate replicates of three samples. Most of the values in the scatter plots corelate very well but a few genes were observed to show big differences between replicates. No extreme outlier was observed in the comparison of WT samples. However, two rrf genes having higher counts are slight outliers in scatter plot of WT_1vWT_3 and WT_2vWT_3 (Figure 5.18-A,B and Figure 5.19-A). The gene IDs of these two genes (having same name) are MM1218R_RS20400 and MM1218R_RS20420. Both the genes code for 5S ribosomal RNA. Both the genes significantly downregulate in KOvWT with log2FoldChange of -1.415 and -1.43 respectively with padj values of 0.04. Both genes upregulate significantly in COMvKO with a log2FoldChange of 2.42 and 2.45 respectively with padj values of 0.0002. Given that this is a rRNA, and most of the rRNA is removed during library preparation, this will relate to differences in the efficiency of rRNA depletion in the different libraries, rather than changes in expression. Therefore, rRNA will be excluded from the analysis.

In the comparison of KO samples, the outliers are *nifJ*, MM1218R_RS10425, MM1218R_RS26545, MM1218R_RS26540, MM1218R_RS17925 and MM1218R_RS02375 (Figure 5.19-B and Figure 5.20-A, B). The expression of all the genes is significantly upregulated or downregulated with log2FoldChange less than -1 or greater than 1 but the padj values for all of them are bigger than 0.05 (Table 5.3). Therefore, these genes were automatically filtered out from the list of differentially expressed genes.

The scatter plots between COM samples show a higher level of reproducibility as no significantly off the line genes were shown except one outlier (MM1218R_RS17350) in COM2vCOM3 comparison (Figure 5.22). MM1218R_RS17350 is exclusively upregulated significantly in COMvKO but is not differentially expressed in other two comparisons.

The low count outliers present in the base of the plot were excluded from much of the analysis because of extremely low counts values. Overall, all the nine comparisons of triplicate biological cultures of three strains show that the replicate libraries were reasonably consistent, and that the data is highly reproducible.

Gene name/ ID	Description	log2FoldChange (KOvWT)	padj
nifJ	flavodoxin oxidoreductase	-1.15	0.322
MM1218R-RS10425	hypothetical protein	-1.29	0.28
MM1218R_RS26545	NAD(P)/FAD-dependent oxidoreductase	4.90	NA
MM1218R_RS26540	wax ester/triacylglycerol synthase family O-acyltransferase	3.78	NA
MM1218R_RS17925	alkane 1-monooxygenase	3.00	0.06
MM1218R_RS02375	acyl-CoA dehydrogenase	2.11	0.06

Table 5.3 The log2FoldChange and padj values of a few outliers in the scatter plots of KO samples.

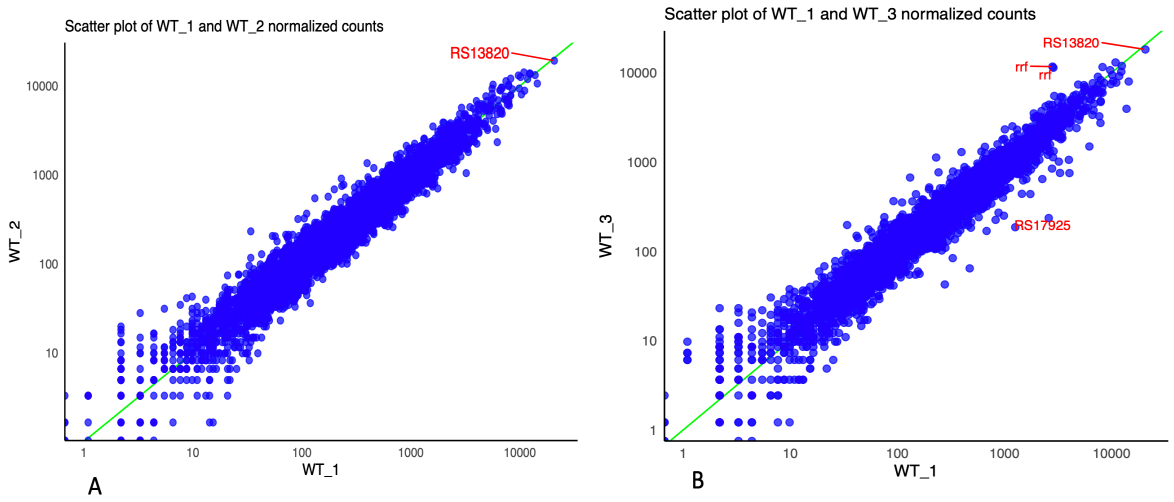


Figure 5.18 Scatter plot of biological replicates WT_1vWT_2 (A) and WT_1vWT_3 (B).

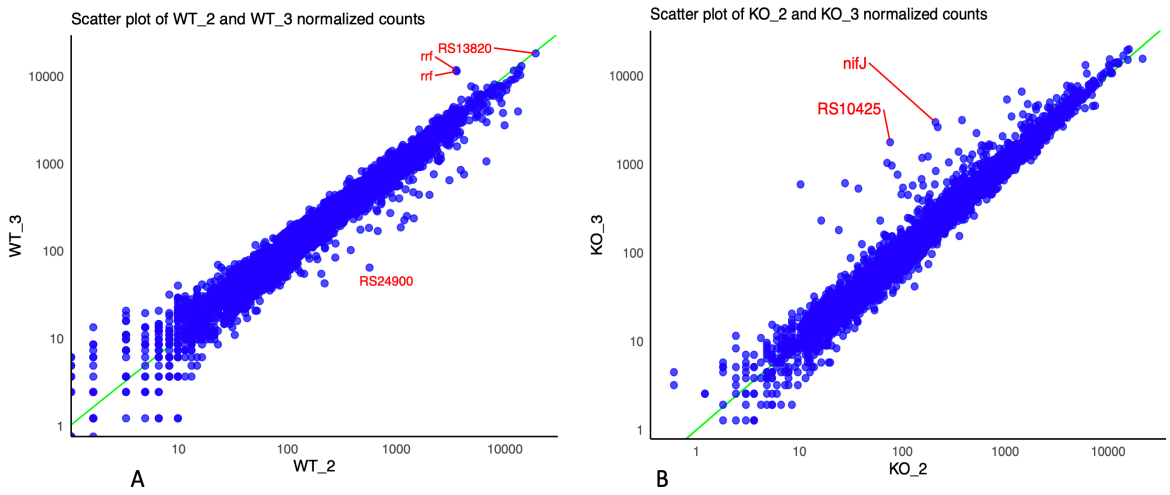


Figure 5.19 Scatter plot between WT_2vWT_3 (A) and KO_2vKO_3 (B).

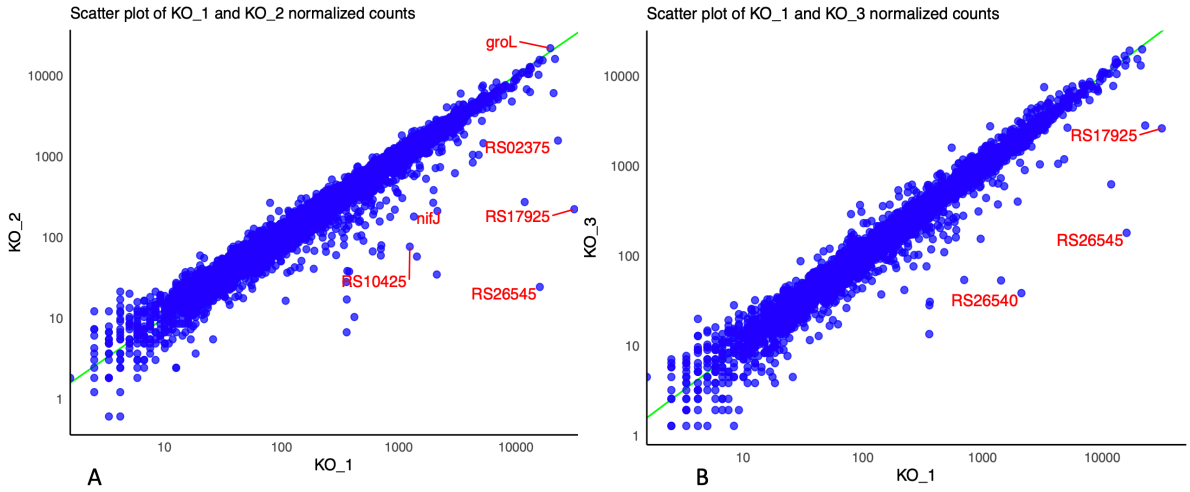


Figure 5.20 Scatter plot between KO_1vKO_2 (A) and KO_1vKO_3 (B).

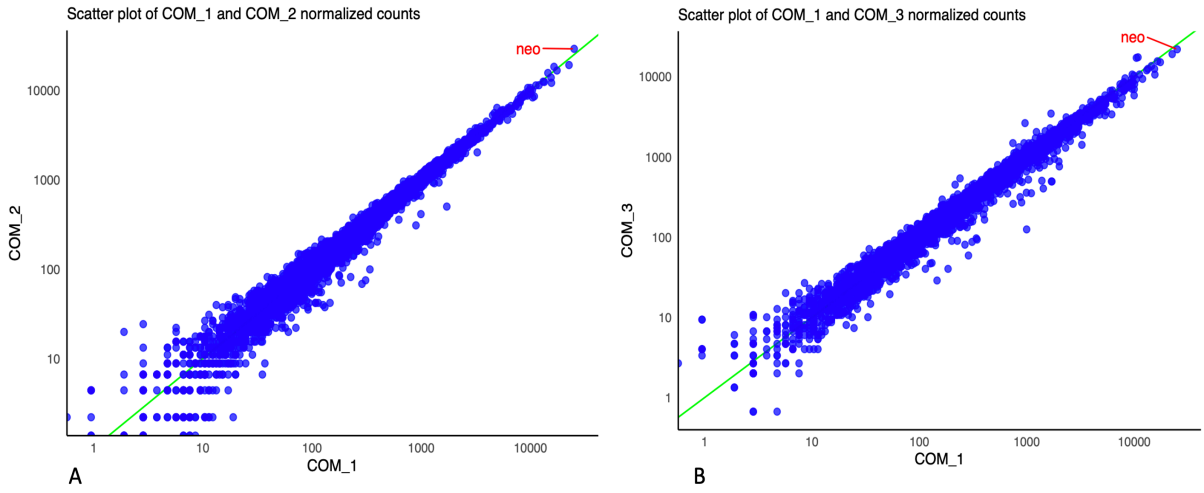


Figure 5.21 Scatter plot between COM_1vCOM_2 (A) and COM_1vCOM_3 (B).

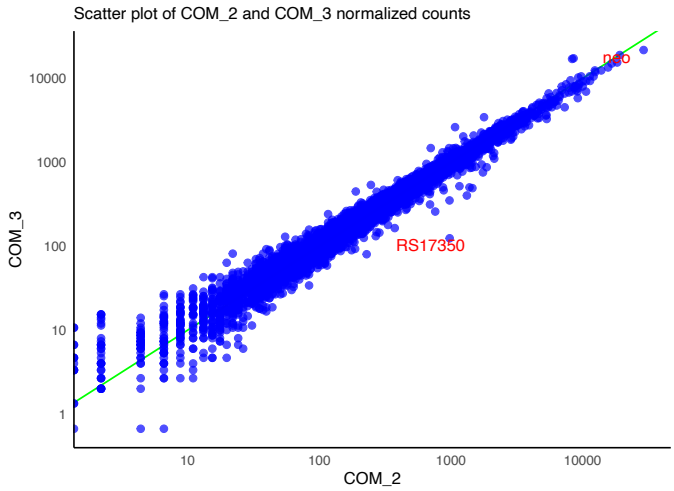


Figure 5.22 Scatter plot between COM_2vCOM_3. Created in R studio using ggplot2 package.

5.3.4.2 Heatmap of normalized counts

To visualize the gene expression across all the nine samples, the normalized counts of all the genes in the KOvWT and COMvKO comparison having log2FoldChange values less than -0.585 or more than 0.585 and padj values less than 0.05 were selected to create the clustered heatmap. The heatmap of all the 2288 genes which fulfilled this criterion is shown in the Figure 5.23. The map can be divided in two parts. The upper part shows the middle 3 columns of KO samples presenting lower counts of gene clusters compared to WT samples. The same gene clusters show higher counts in COM samples nearly similar to the WT samples. The lower part of the heatmap shows upregulation in the expression of genes in KO samples which downregulate in COM samples back to the level of WT samples. Except WT_1 sample, all other samples show high reproducibility in read counts.

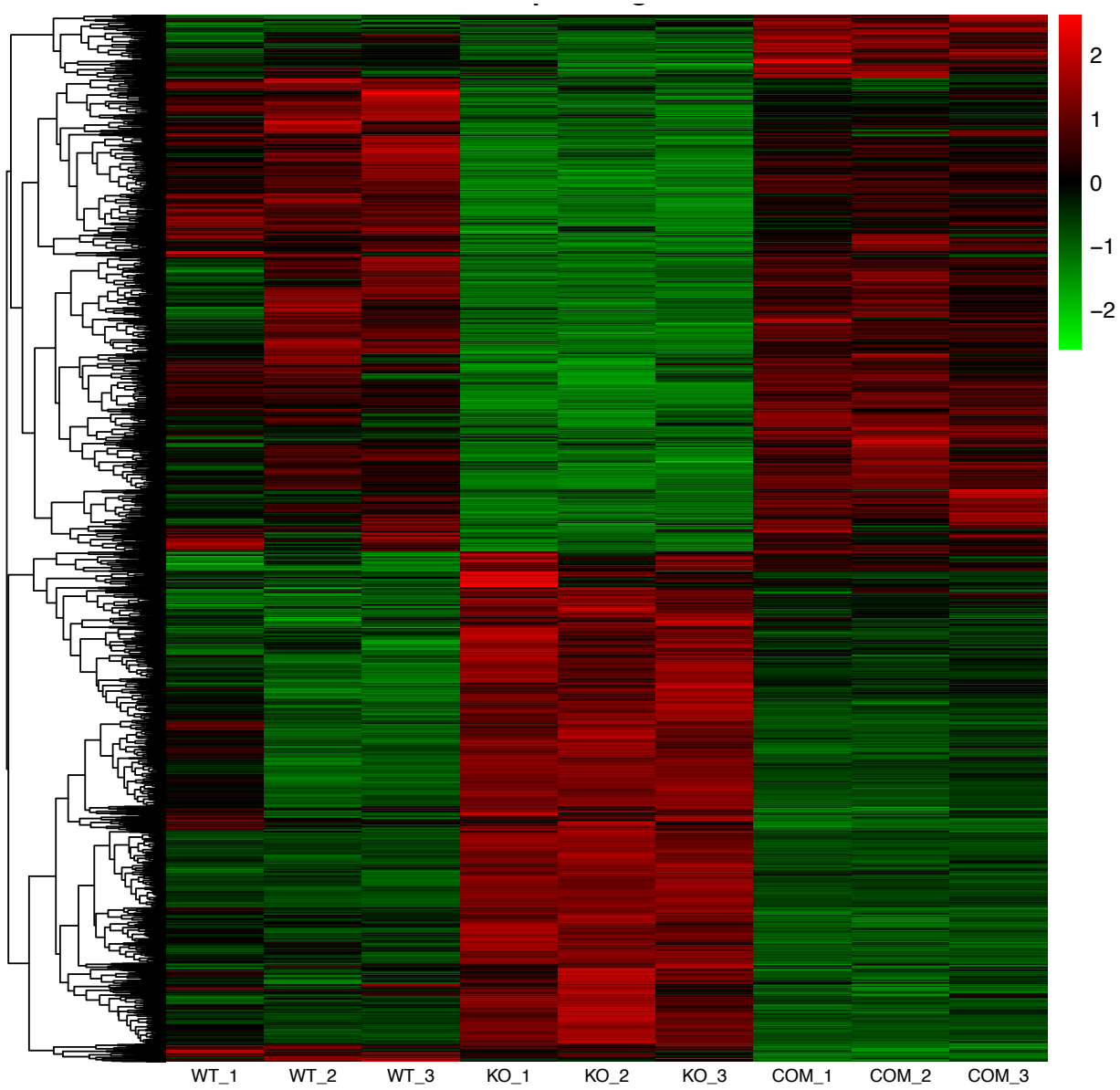


Figure 5.23 Heatmap of normalized counts of triplicate samples of WT, KO, and COM strains. Only 2288 genes having log2FoldChange less than -0.585 and more than 0.585 were selected with padj value cut off <0.05.

5.3.4.3 Scatter plots of log2FoldChange values of pairwise comparison

The log2FoldChange values of pairwise comparisons were also plotted against each other. The scatter plot between the log2FoldChange data of KOvWT and COMvKO demonstrated very reasonable corelation between the two sets of data. Genes that go up in one data set go down in the other and vice versa. Many genes which significantly upregulate in one comparison demonstrate significant down regulation in the other comparison (and vice versa) as most of the points are close to the regression line (Figure 5.24).

The extreme outliers off the line are labelled. The log2FoldChange and padj values of a few outliers are given in Table 5.4. The table shows that the outliers which show significant up or down regulation in KOvWT are not significantly expressed in the other comparison COMvKO either because of high padj value or due to number of normalised counts less than 10. As we are most interested in those genes which are common between the two comparisons i.e., genes significantly upregulate in KOvWT but significantly downregulate in COMvKO and vice versa therefore, these genes will be filtered out from the analysis. Overall, the scatter plot between the two comparisons (KOvWT and COMvKO) show that the results of the RNA-seq data are promising.

Gene name/ ID	Description	log2FC (KOvWT)	log2FC (COMvKO)	padj (KOvWT)	padj (COMvKO)
MM1218R_RS06280	alkane 1-monooxygenase	4.26	-0.93	0.001	0.5
MM1218R-RS19245	transglycosylase family protein	2.58	-0.96	2.72e-09	0.02
MM1218R_RS26545	NAD(P)/FAD-dependent oxidoreductase	4.90	-3.19	NA	NA
MM1218R_RS26540	wax ester/ O-acyltransferase	3.78	-0.89	NA	NA
MM1218R_RS19170	cytochrome P450	2.67	-0.83	0.004	0.41
MM1218R_RS19910	PE family protein	2.65	-1.61	0.0006	0.02
MM1218R_RS07360	hypothetical protein	-4.56	2.86	0.02	0.21
MM1218R_RS07550	hypothetical protein	-3.83	1.63	0.089	0.54

Table 5.4 The log2FoldChange and padj values of a few outliers in the scatter plots 5.22.

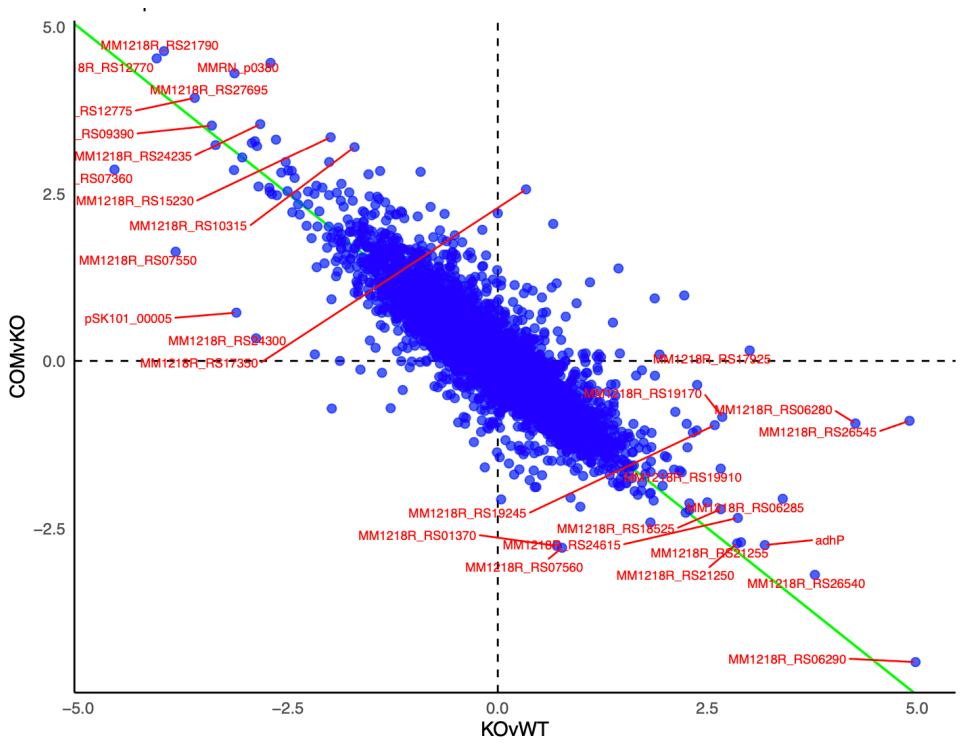


Figure 5.24 Scatter plot between log2FoldChange values of all the genes in KOvWT and COMvKO comparisons. Created in R studio using ggplot2 package.

The scatter plot between the log2FoldChange values of KOvWT and COMvWT, as expected, did not show reasonable corelation. All the points are gathered in the middle of the plot (Figure 5.25 and Figure 5.26). The extreme outliers labelled in the plot show that many of the genes have very high padj values and some are the same genes which were extreme outliers in the previous comparison (Table 5.4) such as MM1218R_RS26540 and MM1218R_RS06280. The outliers will not be included in the analysis.

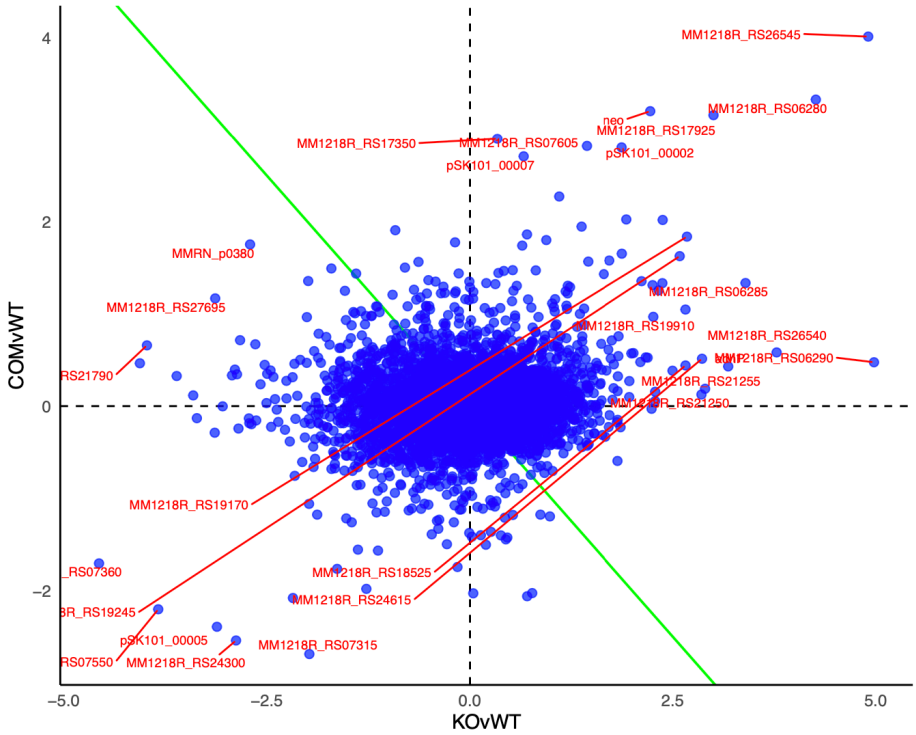


Figure 5.25 Scatter plot between log2FoldChange values of all the genes in KOvWT and COMvWT comparisons.
Created in R studio using ggplot2 package.

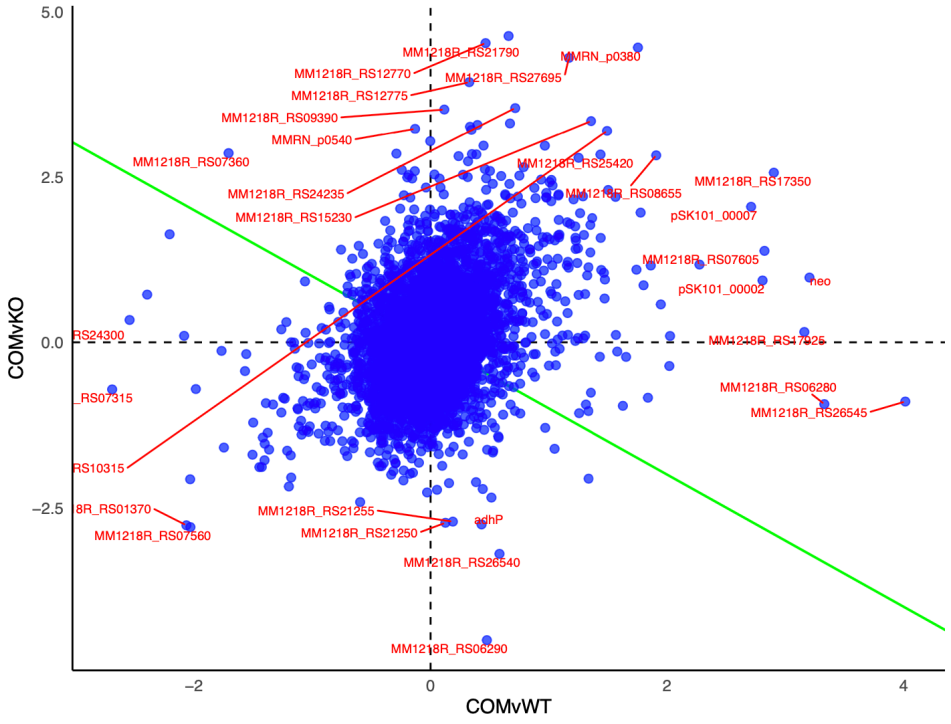


Figure 5.26 Scatter plot between log2FoldChange values of all the genes in COMvWT and COMvKO comparisons.
Created in R studio using ggplot2 package.

5.3.4.4 Differentially expressed genes

To reduce the noise, all the genes having normalised counts less than 10 were removed from the analysis. Two sets of lists of significantly differentially expressed genes were created from the three DESeq2 comparisons datasets. One set contained all the genes having log2FoldChange values less than/equal to -1 and greater than/equal to 1 with padj values less than 0.05. The second set contained all the genes having log2FoldChange values between (-1 to -0.585) and (1 to 0.585) with padj values less than 0.01. The number of genes in each list are given in Table 5.5.

Log2FoldChange < = -1 for downregulated genes, >= 1 for upregulated genes, padj <0.05					
KO v WT		COM v KO		COM v WT	
Up	Down	Up	Down	Up	Down
285	425	480	323	41	23
Log2FoldChange between (-1, -0.585) for downregulated genes (1, 0.585) for upregulated genes, padj <0.01					
KO v WT		COM v KO		COM v WT	
Up	Down	Up	Down	Up	Down
529	354	420	556	28	22

Table 5.5 Two sets of differentially expressed genes sorted by log2FoldChange and padj values of the three comparisons of DESeq2 analysis.

The top 75 genes having the highest/lowest values of log2FoldChange and lowest padj values from the KOvWT comparison were selected to create a clustered heatmap using the normalized counts of those genes. The log2FoldChange values range from 3.18 to 1.0 for upregulated genes and from -1.8 to -3.6 for downregulated genes. 64 genes have extremely low padj values ranging from 6.63e-121 to 6.22e-07 and remaining genes have padj values less than 0.01.

The heatmap demonstrated that from the list of 75 all the genes which upregulate in KO, downregulate in COM and vice versa (Figure 5.27). Interestingly, among upregulated genes (in KO), three molecular chaperones i.e., Cpn60.2 (GroL), DnaJ and DnaK and a co-chaperonin *cpn10* (GroES) were included. Other CDS upregulated in KO were nine genes coding for PE and PPE family proteins, two DUF family proteins, five transporter proteins and various enzymes such as fatty acid desaturase,

glycosyltransferase, and alcohol dehydrogenase. The heatmap also demonstrated that the top downregulated genes in KO encoding for various enzymes upregulated in complemented strain back to the level nearly equivalent to that of WT samples.



Figure 5.27 Heatmap of normalized counts of triplicate samples of WT, KO, and COM strains. Only top 75 genes having log2FoldChange less than -1 and more than 1 were selected with padj value cut off <0.01. Chaperone genes are highlighted in yellow, PE and PPE family proteins in green, genes coding for transporter protein in blue and DUF domain containing proteins in purple. Created in R studio using pheatmap package.

5.3.4.5 Venn diagrams – common and exclusive up and down regulated genes.

The two sets of significantly up and down regulated genes listed in Table 5.5 were investigated further to find the common and exclusive genes among different groups. In first set of log2FoldChange cut off value 1, out of the 285 significantly upregulated genes in KOvWT, 184 genes were found to be downregulated in COMvKO. And out of the 425 genes significantly downregulated in KOvWT, 315 were found to be up regulated in COMvKO (Figure 5.28-A). These two common sets of genes are of higher interest in this analysis. Therefore, these two sets were analysed in detail which is discussed in the next sections.

In second set of data of log2FoldChange cut off between 1 and 0.585, out of 529 gene significantly upregulated in KOvWT, 314 were found to be downregulated in COMvKO comparison. Whereas, out of the 354 genes significantly downregulated in KOvWT comparison, 171 were found to be upregulated in COMvKO (Figure 5.28-B). These two common sets of genes are discussed in next sections. For a quick look of common and exclusive genes in KOvWT and COMvKO comparison see Table 5.6.

The comparison of genes up and downregulating in KOvWT with those of COMvWT, as expected, did not show any common gene upregulating in KOvWT but downregulating in COMvWT in log2FC cut off 1 list; only one gene was in this category in log2FC betweem 1 - 0.585 list. From log2FC 1 list, only 6 genes were found to be commonly downregulating in KOvWT and upregulating in COMvWT (Figure 5.28-C). One of the 6 common genes is PE-PGRS family protein, one is a hypothetical protein, and one is DUF family protein (Table 5.7).

Eleven genes were found upregulating commonly in KOvWT and COMvWT (log2FC 1 cut off). Two of these were *sacB* and hygromycin resistance gene which were present only in the chromosome of the KO and COM strains. And two of them were pSK101-00001 (Kanamycin resistance cassette) and pSK101-00002 (pSK101 backbone gene). Although the pSK101 vector (without *cpn10* and *cpn60.1* genes) was transformed in the WT cells. We don't know the reason why the expression of some plasmid genes goes up in KO or COM strains compared to the WT.

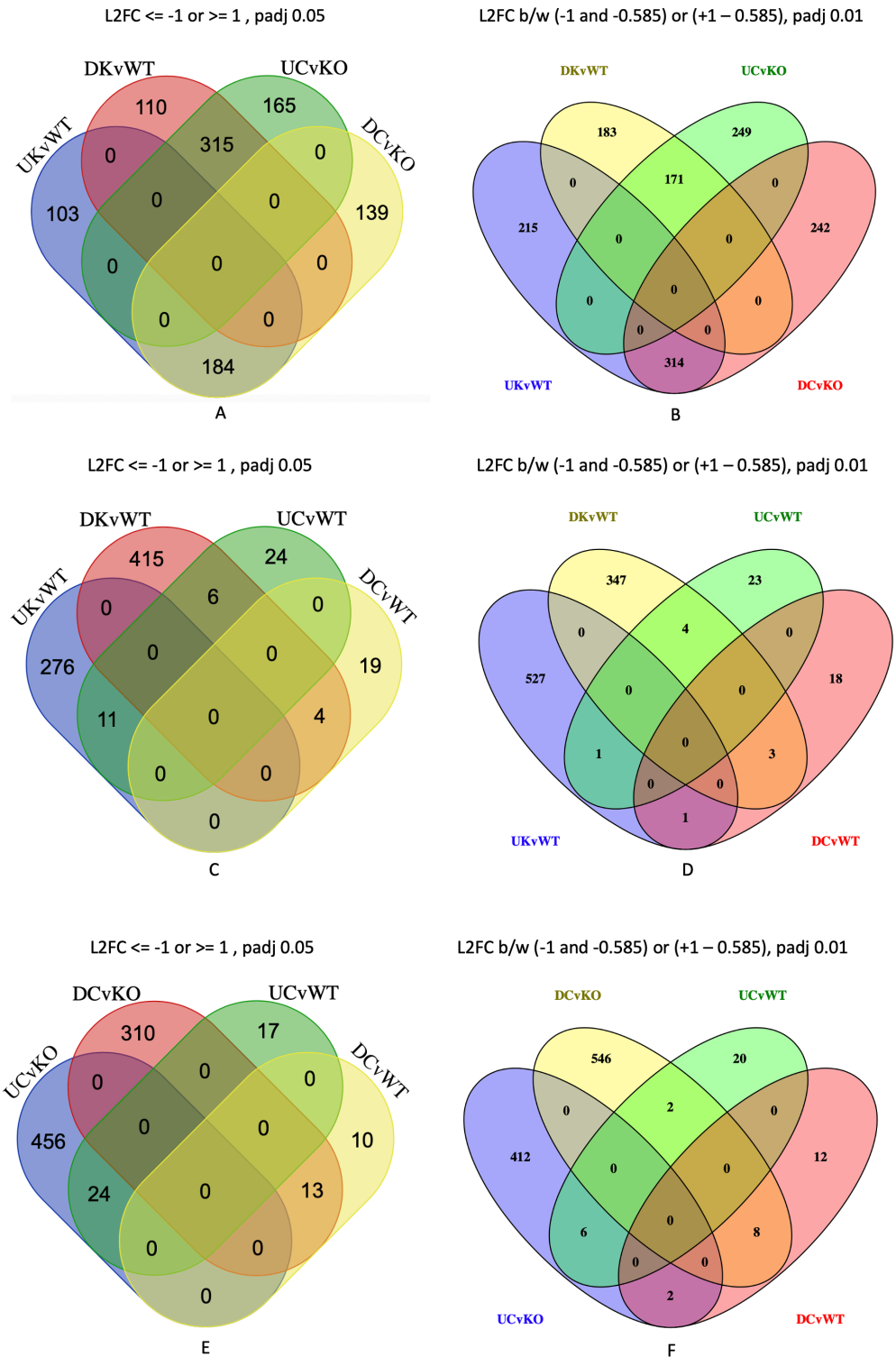


Figure 5.28 Venn diagrams among various combinations of up and down regulated genes showing exclusive and common regulated genes.
Abbreviations: UK_WT (Up in KO v WT), DK_WT (Down in KO v WT), UCvKO (Up in COM v KO), DCvKO (Down in COM v KO), UCvWT (Up in COM vWT), and DCvWT (Down in COM vWT). (Figures created using Venny 2.1.0 – BioinfoGP and https://www.vandepeerlab.org/?q=software.

Common genes	L2FC cut off 1	L2FC b/w 1 & 0.585	Exclusive genes	L2FC cut off 1	L2FC b/w 1 & 0.585
UKvWT_DCvKO	184	314	UKvWT	103	215
			DKvWT	110	183
DKvWT_UCvKO	315	171	UCvKO	165	249
			DCvKO	139	242

Table 5.6 List of common and exclusive genes in KOvWT and COMvKO comparisons.

Gene name/ ID	Description	log2FC (KOvWT)	padj (KOvWT)	log2FC (COMvWT)	padj (COMvWT)
MM1218R_RS15230	DoxX family protein	-1.99	1.93e-23	1.357	1.48e-13
MM1218R_RS10315	tRNA-Cys	-1.706	0.004	1.49	0.028
MM1218R_RS09785	PE-PGRS family protein	-1.53	6.94e-08	1.25	6.50e-06
MM1218R_RS25420	hypothetical protein	-1.40	2.68e-05	1.43	5.84e-06
MM1218R_RS07890	NAD(P)/FAD-dependent oxidoreductase	-1.34	8.61e-07	1.03	0.001
MM1218R_RS28410	DUF6131 family protein	-1.11	1.25e-06	1.01	0.001

Table 5.7 Six genes common downregulating in KOvWT and upregulating in COMvWT comparisons.

(log2FC cut off 1, padj <0.05).

Among the comparisons of COMvKO and COMvWT, 24 genes were commonly found to be upregulated in both comparisons (log2FC 1 cut off) and 6 genes in this category were from log2FC b/w 1 – 0.585 list. Whereas 13 (L2FC 1) and 8 (L2FC 1 – 0.585) genes were found to be commonly downregulated (Figure 5.28-E & F).

5.3.4.6 Scatter plots of differentially expressed common genes

The common genes upregulated in knockout but down regulated in complement (UKvWT_DCvKO) and vice versa (DKvWT_UCvKO) were of higher interest in this analysis because our focus was to find the direct or indirect effect of deletion of *cpn60.1* on the expression of other genes. The 184 genes found commonly upregulating in KOvWT but downregulating in COMvKO and 315 genes commonly downregulating in KOvWT but upregulating in COMvKO with log2FC cut off 1 and padj values less than 0.05 were analysed further.

Out of the 184 common UKvWT_DCvKO genes, 92 genes had padj values less than 10e-10 and 85 genes had padj values less than 0.01 in both the individual comparisons. Only two had padj value more than 0.01 (but < 0.05) in both the comparisons and remaining five genes had padj values less than 0.01 in one comparison but > 0.01 < 0.05 in the other one (Figure 5.29). In case of 315 common genes DKvWT_UCvKO 120 were noticed to had padj values < 10e-10 and 178 < 0.01 (Figure 5.30). This showed that the data was very reliable.

Because we are interested in the change/effect caused due to the absence of Cpn60.1 in the cells. Technically the amount of up or downregulation in a gene affected directly or indirectly due to the absence of Cpn60.1 should return with same magnitude when Cpn60.1 is complemented in the cells. That means, ideally, if a gene was upregulated with log2FC of +4 in KOvWT then in COMvKO the same gene should downregulate with -4 log2FC. In that case the ratio of the log2FC values will be -1. Closer the ratio of log2FC values to -1, more reliable the data and it reinforces the argument that the effect was caused by the absence of Cpn60.1. The extreme outliers could be excluded from the analysis as they probably be just noise.

The scatter plots of 184 (UKvWT_DCvKO) and 315 (DKvWT_UCvKO) common genes showing the colour coded genes according to the ratios of log2FC values demonstrated that most of the genes had the ratio of log2FC values close to -1 (Figure 5.31 & Figure 5.32). A few extreme outliers are labelled. Most of the these are either PE family proteins or hypothetical proteins (Table 5.8).

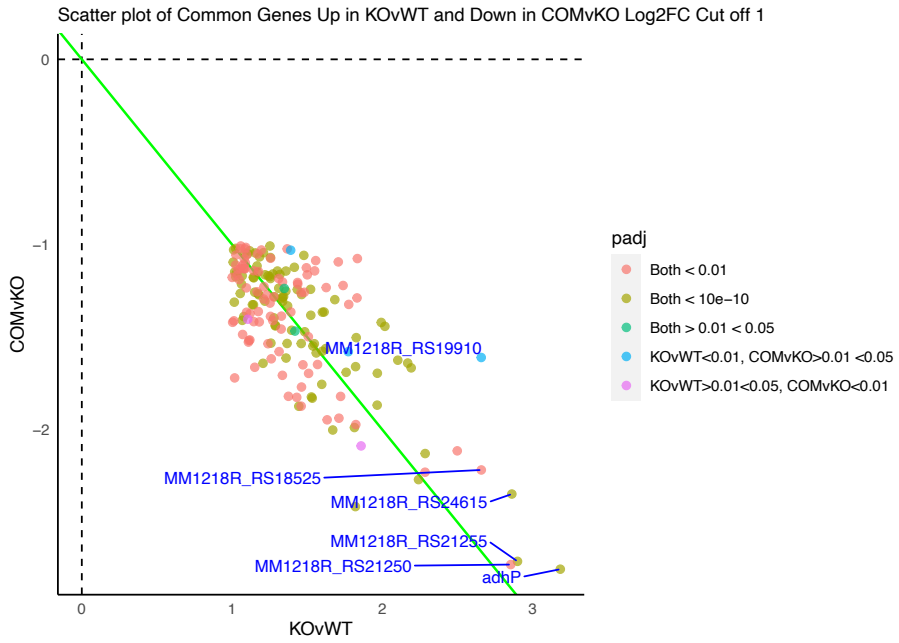


Figure 5.29 Scatter plot of common genes up in KOvWT but down in COMvKO. UKvWT_DCvKO colour coded according to the padj values. Created in R studio using ggplot2 package.

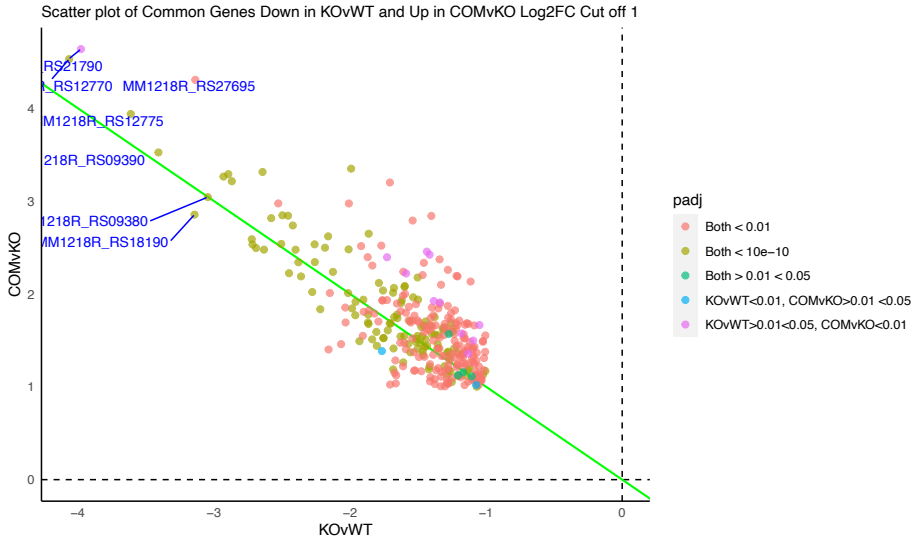


Figure 5.30 Scatter plot of common genes down in KOvWT but up in COMvKO. DKvWT_UCvKO colour coded according to the padj values. Created in R studio using ggplot2 package.

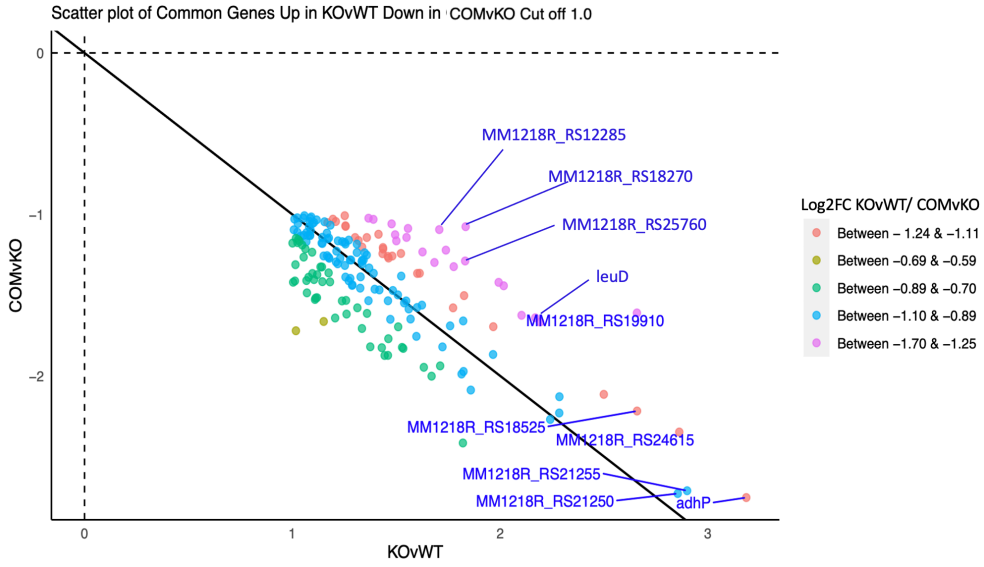


Figure 5.31 Scatter plot of common genes down in KOvWT but up in COMvKO. DKvWT_UCvKO colour coded according to the values obtained by dividing of log2FC values of KOvWT by COMvKO. (R studio using ggplot2 package).

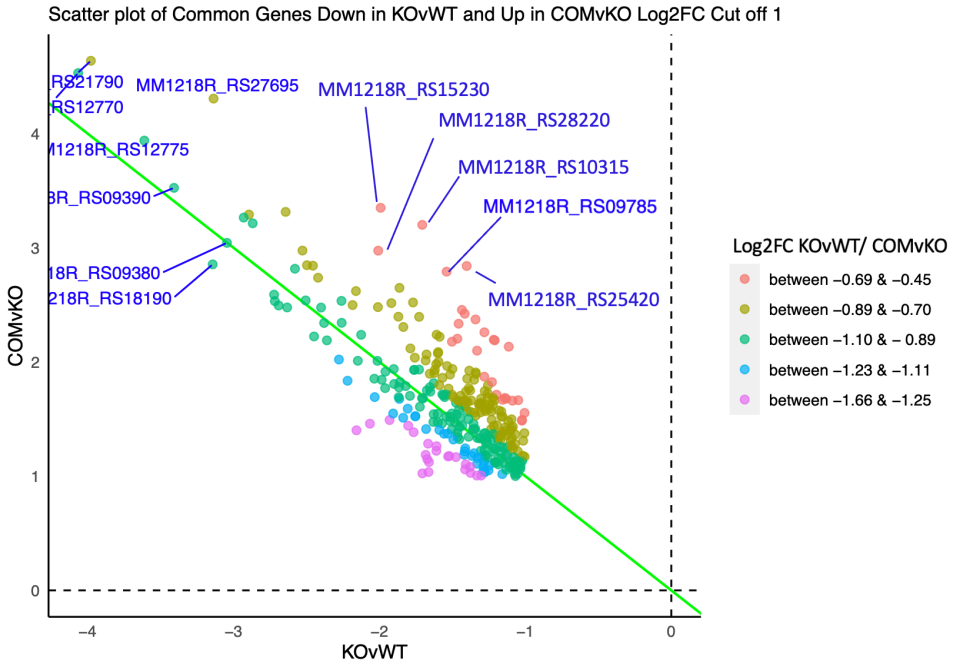


Figure 5.32 Scatter plot of common genes down in KOvWT but up in COMvKO. DKvWT_UCvKO colour coded according to the values obtained by dividing of log2FC values of KOvWT by COMvKO. Created in R studio using ggplot2 package.

The analysis of padj values and log2FC values of common genes (log2FoldChange cut off 1, padj <0.05) upregulating in KOvWT but downregulating in COMvKO and vice versa showed that the data was very reliable. Only a few genes had padj values greater

than 0.01. And the expression of genes significantly going up in KO samples downregulated to nearly the same level in the complemented samples.

Gene name/ ID	Description	log2FC (KOvWT)	padj (KOvWT)	log2FC (COMvWT)	padj (COMvWT)
MM1218R_RS12285	PE domain-containing protein	1.8339174	2.55e-12	-1.076009	2.57e-05
MM1218R_RS18270	PE family protein	1.7070292	1.46e-10	-1.094048	9.37e-06
MM1218R_RS25760	PE family protein	1.8315530	0.000366	-1.2885522	0.005178
MM1218R_RS19910	PE family protein	2.6581356	0.000636	-1.61032	0.014479
MM1218R_RS15230	DoxX family protein	-1.991646	1.93e-23	3.3488684	4.00e-72
MM1218R_RS28220	hypothetical protein	-2.0082158	0.000185	2.9737800	1.25e-08
MM1218R_RS10315	tRNA-Cys	-1.7065060	0.004730	3.1997852	1.61e-09
MM1218R_RS09785	PE-PGRS family protein	-1.539494	6.94e-08	2.7901026	1.79e-27
MM1218R_RS25420	hypothetical protein	-1.4020041	2.68e-05	2.8397433	2.83e-22

Table 5.8 Table showing the names, log2FoldChange, and padj values of the outliers in figure 5.31 and 5.32.

5.3.4.7Volcano plots

To visualize both significant and non-significant genes with respect to log2FC and padj values volcano plots of the three DESeq2 analysis comparisons were created. The volcano plot of KOvWT comparison between log2FC values on x-axis and statistical significance -log10 padj values on y-axis is shown in Figure 5.33. Significantly up and down regulated genes having the log2FC values more than or equal to 1 and padj values < 0.01 are displayed in red towards the right and left side of the plot respectively. Many genes in blue have padj values less than 0.01 but because of not fulfilling the criteria of log2FC cut off they are not included in the significantly differentially expressed genes in the plot. However, as it was discussed in section 5.3.4.4 that 529 (up) and 354 (down) more genes could be listed in the category of significantly expressed genes if the criteria of log2FC is decreased from 1 to 0.585.

Among the significantly upregulated genes displayed in the plot, large number of genes showed -log10 padj values more than 10 which indicated extremely low padj values and that the chances of the results to happen by chance were negligible. Some important genes are labelled. Some genes encoding for molecular chaperones such as *groEL* (encoding Cpn60.2), *groES* (encoding Cpn10), *dnaJ*, *dnaK* are shown among the significantly upregulated genes. 13 top listed significantly differentially expressed genes in KOvWT comparison are not shown in the plot 5.31 because of their extremely low padj values (and in turn very high -log10 padj) out of the limit of the y-axis.

The volcano plot of COMvKO comparison is shown in Figure 5.34. Colour coding, x, y - limits and criteria for significantly differentially expressed genes is same as described for KOvWT comparison. Genes encoding molecular chaperones such as *groL, groES, dnaJ*, and *dnaK* are labelled among significantly downregulated genes which demonstrated that the same genes for molecular chaperones which upregulated in KOvWT comparison were downregulated in COMvKO comparison and that the padj values of those genes were nearly equivalent in both comparisons. The comparison of significantly downregulated genes labelled in KOvWT volcano plot showed that many genes were present in the significantly upregulated labels in COMvKO such as MM1219R_RS22920, MM1219R_RS09393, MM1219R_RS24200, MM1219R_RS01030, and *hypF*.

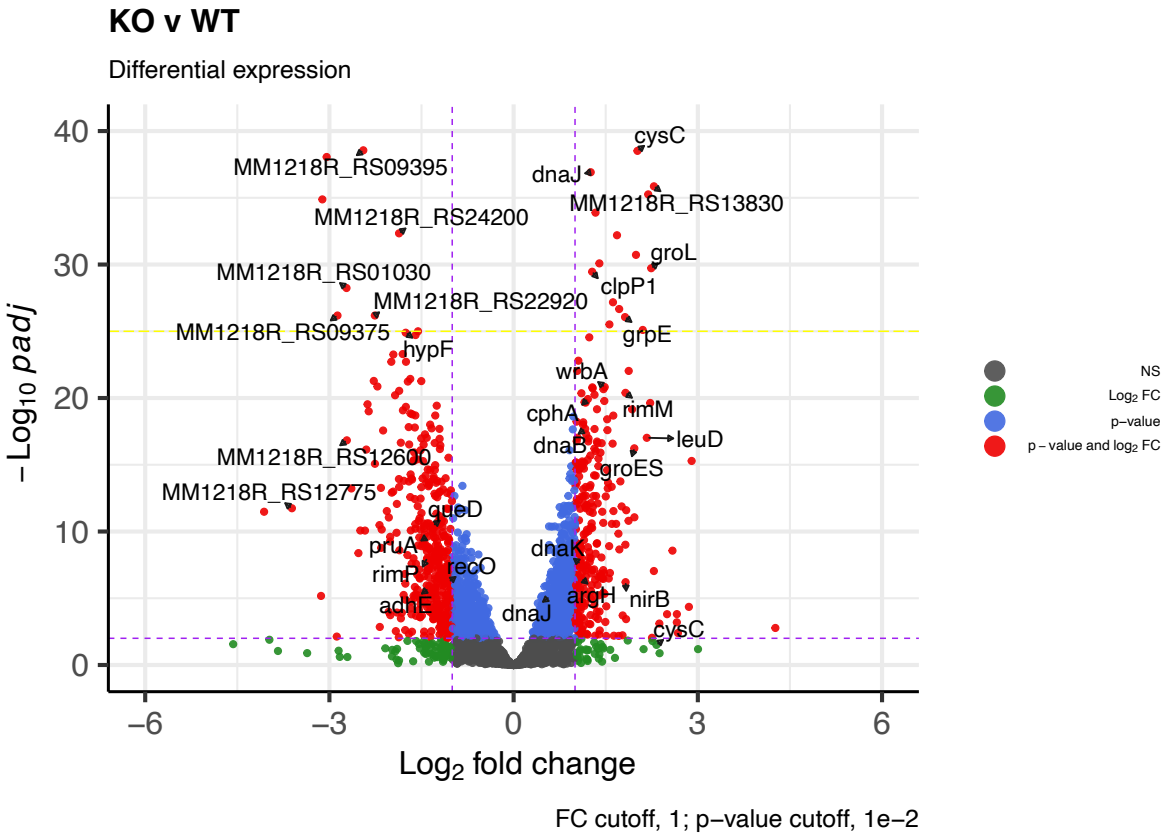


Figure 5.33 Volcano plot of KOvWT DESeq2 analysis created in R studio using EnhancedVolcano package.
Red dots indicate differentially expressed genes falling under the two criteria of L2FC cut off 1 and padj cut off 0.01. Blue dots represent genes having padj value less than 0.01 and green dots indicate genes having log2FC values greater than 1 or less than -1 but padj values bigger than 0.01. Black dots are non-significantly expressed genes.

The volcano plot of COMvWT comparison did not show many significantly differentially expressed genes (Figure 5.35). This was expected because the complemented strain should technically behave like WT strain and that the effects of the deletion of *cpn60.1* technically should vanish in the presence of plasmid copy of *cpn60.1* and the cells should return to the normal regulation state. A few differences observed could be because on chromosome only one copy of *cpn60.1* is present in WT strain whereas in COM strain the pSK101 plasmid express more than one copies of the plasmid (and hence *cpn60.1* too) in the cell. The higher levels of Cpn60.1 might also affect the expression of other genes in the cell.

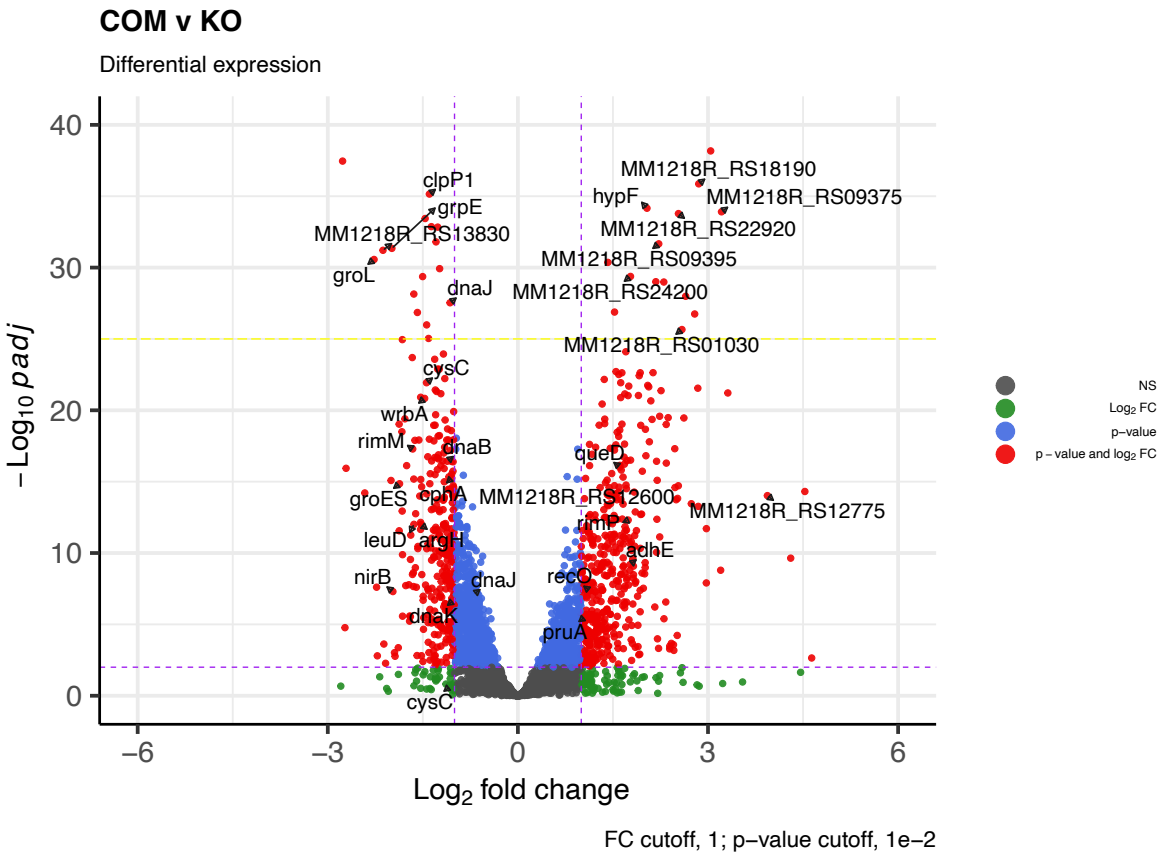


Figure 5.34 Volcano plot of COMvKO DESeq2 analysis created in R studio using EnhancedVolcano package.

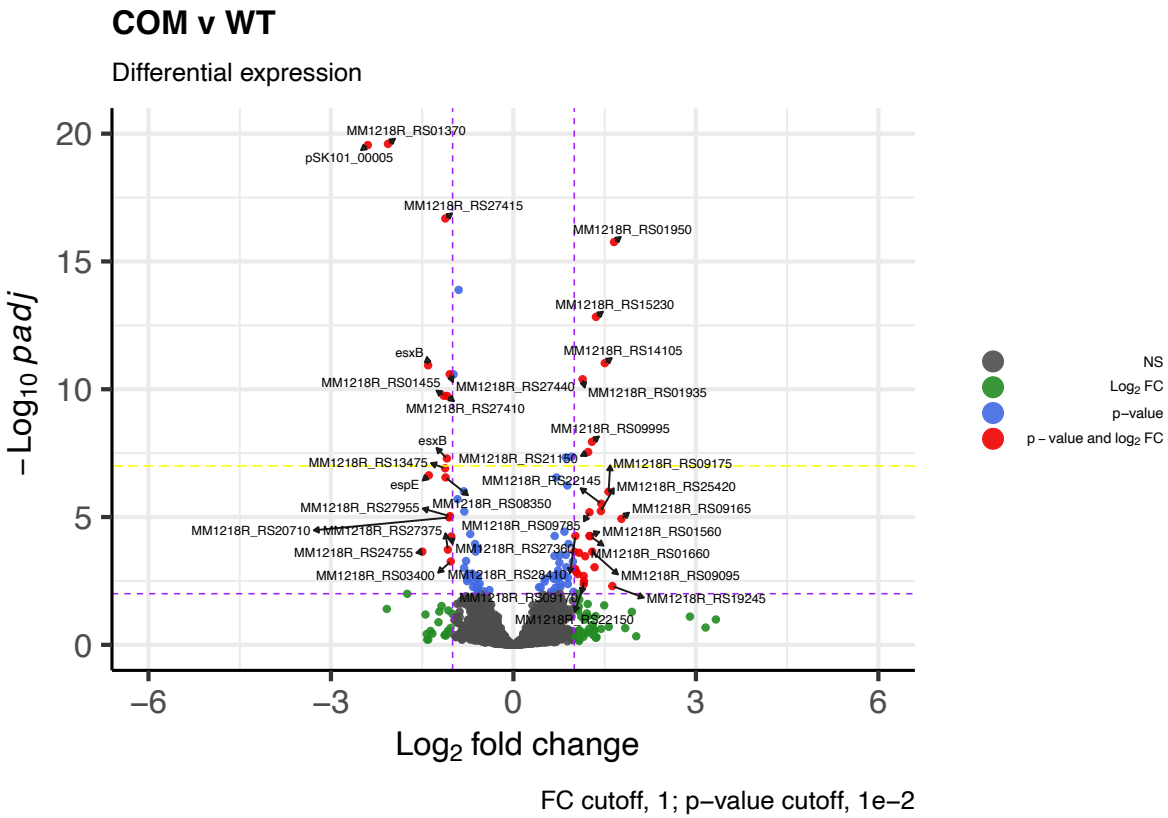


Figure 5.35 Volcano plot of COMvWT DESeq2 analysis created in R studio using EnhancedVolcano package.

Different methods of data analysis discussed above have strongly favoured the hypothesis that the deletion of *cpn60.1* has a large impact on the transcriptomic profile of the *M. marinum*. And that the effect caused is reversed to a great extent when loss of chromosomal *cpn60.1* is compensated by plasmid expressed copy of *cpn60.1*.

Further analysis was done on the lists of common genes UKvWT_DCvKO and DKvWT_UCvKO. To investigate the category of genes significantly up or down regulating and the biological or cellular pathways they belong to, gene ontology pathway analysis was performed which is discussed in the following section.

5.3.5 Gene ontology enrichment analysis

Different gene ontology pathway analysis programs support genomes of a restricted list of organisms. Most of such websites or programmes are designed for pathway analysis or gene set enrichment analysis of eukaryotic organisms such as human, mice, drosophila, and zebrafish. A couple of online software were found which were able to support either the genome of *M. tuberculosis* H37Rv or *M. marinum* M strain. The *M. marinum* 1218R genome is not completely annotated yet. The annotation of many CDS sequences is missing and many duplicated gene names are present in the NCBI annotation document. Therefore, to perform the gene ontology pathway analysis of the RNA-seq data, the Roary pangenome analysis was performed between 1218R and M strains to find out the homologous genes between 1218R and M strain. The Roary analysis was also performed between 1218R and *M. tuberculosis* H37Rv. The homologous genes of M strain and H37Rv were then used for pathway analysis. The details of Roary analysis are discussed in chapter 3, section 3.3.1.8. An online tool http://bioinformatics.sdstate.edu/ was used to perform gene ontology enrichment analysis of both the *M. marinum* M and Mtb H37Rv homologues genes with FDR cut off 0.05.

5.3.5.1 Pathway analysis of common genes UKvWT_DCvKO

5.3.5.1.1 Analysis using *M. marinum* M homologous genes – log2FC cut off 1, padj value < 0.01

From the list of 184 common genes with log2FC cut off 1 and padj value cut off 0.05 the 7 genes having padj value more than 0.01 were excluded from the analysis. By performing Roary pangenome analysis, out of 177 genes, 170 genes were found to have homologous genes (at least 70% homology) in M strain. The *M. marinum* M homologs of 170 genes obtained by Roary analysis were used for the gene ontology enrichment analysis. As described below, three interesting sets of pathways associated with the genes under question were found.

- **Stress response, protein folding and chaperone activity**

Many genes in the list were found to be involved in stress response, protein folding or chaperone activity. Stress response pathway was observed to have the highest fold

enrichment (Figure 5.36, Table 5.10). The genes involved in stress response in this analysis were *dnaK*, grpE, dnaJ and hspR. Two genes encoding proteins primarily involved in protein folding were *groES* (*cpn10*) and *groL* (*cpn60.2*). Gene encoding molecular chaperones, co-chaperones or regulatory protein for chaperones were *dnaK, grpE, dnaJ, hspR, cpn60.2, cpn10*, *tig* and *clpP1.* In KOvWT comparison, among molecular chaperones *cpn60.2* demonstrated the highest log2FoldChange of 2.24, *cpn10* 1.96, *dnaJ* 1.25, *dnaK* 1.09, *hspR* 1.52, *grpE* 1.81, and tig was upregulated by 1.15 log2FC. These chaperone genes showed downregulation in COMvKO comparison with the log2FC of -2.26 for *cpn60.2,* -1.86 for *cpn10*, -1.07 for *dnaJ,* -1.01 for *dnaK*, -1.38 for *hspR,* -1.98, and -1.04 for *tig*. All the genes had very low padj values. The ratio of log2FC values of upregulation and downregulation of these chaperone genes was calculated to be very close to 1 (Table 5.9). This showed that the absence of cpn60.1 caused an increase in the expression of other chaperones associated with protein folding and stress response and that the effect was reversed to the same mangnitude in the complemented samples. The low padj values reinforced the realiability of the results.

Clp proteases play essential proteolytic role in protein turnover in the cells by degrading dysfunctional, misfolded, and short-lived proteins. They are also considered as a promising drug target against *M. tuberculosis*. Out of the two Clp proteases in *M. tuberculosis*, Clp P1 is considered to be essential for cell survival which could be because of its essential role in degradation of some specific substrates (d'Andrea et al., 2022).

- **Cell wall, cell division, and peptidoglycan synthesis pathway**

The pathway analysis demonstrated that four *mur* family genes *murD, murX, murF*, and *murE* were associated with cell cycle, cell division, cell wall biogenesis/degradation, peptidoglycan synthesis and maintenance of cell shape and *ftsX* gene encoding a cell division protein FtsX was also involved in these pathways. *murD, murX, murF*, and *murE* demonstrated upregulation in KOvWT by a log2FoldChange of 1.119, 1.15, 1.116, and 1.24. These genes downregulated in COMvKO comparison with log2FoldChange -1.37, -1.25, -1.03, and -1.16 for *murD, murX, murF,* and *murE* respectively. ftsX showed log2FC of 1.44 and 3.31e-12 padj value in KOvWT whereas in COMvKO its log2FC was -1.87 and 9.34e-20 padj value (Table 5.11). *mur* family genes play an essential role in

the biogenesis of peptidoglycan (or murein) which is an essential part of bacterial cell wall (Munshi et al., 2013). The hierarchical clustering tree indicates that all these pathways are interconnected with each other (Figure 5.37).

- **Polyketide synthase pathways**

A third set of interconnected pathways observed by GO pathway analysis were polyketide synthase pathways. Several genes such as *fadD30, pks6, pks5, pks12, mbtB, mbtF, mbtC*, *drrA, ppsD,* and some associated membrane proteins in the list of log2FC 1 cut off padj<0.01 were associated with polyketide synthase pathways (Table 5.10). The log2FC and padj values of these genes are given in Table 5.12.

Polyketides are a large and diverse group of natural products, which are basically secondary metabolites that have wide range of applications from medicines to toxicity. Polyketide synthases (PKSs) are a large family of multidomain and multifunctional enzymes that are involved in the biosynthesis of diverse polyketides. In *M. tuberculosis* several polyketide synthase genes are found to be associated with exclusive biosynthetic pathways involved in the production of various unique complex lipids and conjugated glycolipids that are either required for virulence or form major components of complex cell wall of mycobacteria (Quadri, 2014). These pathways are of immense importance for biogenesis of mycobacterial cell wall. The full list of common genes UKvWT_DCvKO along with the log2FC and padj values and homologous genes of M strain and Mtb strain is given in supplementary data.

There was possibility that other genes involved in these pathways were also significantly differentially expressed but their log2FC values were slightly lower than 1. Therefore, I looked from the list of genes which upregulated in KOvWT with log2FC values between 1 – 0.585 and padj values < 0.01 if there were more significantly upregulated genes associated with these pathways.

Sr No.	ID	gene_name	description	log2FC_KOvWT	padj_KOvWT	log2FC_COMvKO	padj_COMvK	M strain hoomolog	M gene_name	Rv homolog	Rv gene_name
1) Stress response, protein folding and chaperone activity											
1	MM1218R_RS03575	groL	chaperonin GroEL, Prevents misfolding and promotes the refolding and proper assembly of unfolded polypeptides generated under stress conditions.	2.240840417	1.89E-30	-2.269858838	2.71E-31	MMAR_0759	groEL2	Rv0440	groEL2
2	MM1218R_RS05440	groES	co-chaperone GroES, binds to Cpn60 in the presence of Mg-ATP and suppresses the ATPase activity of the latter	1.965859343	5.86E-17	-1.867575292	1.40E-15	MMAR_1125	groES	Rv3418c	groES
3	MM1218R_RS02970	dnaJ	Participates actively in the response to hyperosmotic and heat shock by preventing the aggregation of stress-denatured proteins, in association with DnaK and GrpE. It is the nucleotide exchange factor for DnaK and may function as a thermosensor.	1.255551749	1.22E-37	-1.070368011	2.93E-28	MMAR_0639	dnaJ	Rv0352	dnaJ1
4	MM1218R_RS02960	dnaK	Participates actively in the response to hyperosmotic and heat shock by preventing the aggregation of stress-denatured proteins and by disaggregating proteins, also in an autonomous, DnaK-independent fashion.	1.091026703	2.63E-08	-1.015666558	2.13E-07	MMAR_0637	dnaK	Rv0350	dnaK
5	MM1218R_RS19345	clpP1	ATP-dependent CLP protease proteolytic subunit.Cleaves peptides in various proteins in a process that requires ATP hydrolysis. Has a chymotrypsin-like activity. Plays a major role in the degradation of misfolded proteins.	1.27882448	3.46E-30	-1.396290633	7.06E-36	MMAR_3808	clpP1	Rv2461c	clpP1
6	MM1218R_RS02965	grpE	Participates actively in the response to hyperosmotic and heat shock by preventing the aggregation of stress-denatured proteins, in association with DnaK. It is the nucleotide exchange factor for DnaK.	1.814313588	8.64E-27	-1.988597607	4.50E-32	MMAR_0638	grpE	Rv0351	grpE
7	MM1218R_RS02975	MM1218R_RS02975	Involved in transcriptional regulation (repression) of heat shock proteins e.g., DnaK, GrpE, DnaJ1. binds to three inverted repeats (IR1-IR3) in the promoter region of the DnaK operon. induction: by heat shock	1.523905197	3.92E-19	-1.382196184	2.78E-16	MMAR_0640	hspR	Rv0353	hspR
8	MM1218R_RS19350	MM1218R_RS19350	Involved in protein export. Acts as a chaperone by maintaining the newly synthesized protein in an open conformation. Functions as a peptidyl-prolyl cis-trans isomerase	1.154788168	2.05E-20	-1.045499331	3.86E-17	MMAR_3809	tig	Rv2462c	tig

Table 5.9 Detail of proteins associated with molecular chaperones, stress response and protein folding pathways.
The log2FC and padj values of KOvWT and COMvKO comparison are given. The homologous genes of M strain and H37Rv was determined by Roary analysis. The gene description was taken from GO pathway analysis.

	GO pathway analysis performed using M strain homologs of common genes up in KOvWT and down in COMvKO with log2FC cut off 1 and padj value <0.01					
2	Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway	Genes
3	0.001337958	4	5	25.67100592	Stress response	dnaK grpE dnaJ hspR
4	0.005196909	3	4	24.06656805	Protein folding	dnaK groES tig
5	0.001427376	4	6	21.39250493	Mur ligase, C-terminal, and glycosylation	murD murX murF murE
6	0.001427376	4	6	21.39250493	Mur ligase, C-terminal	Ligase murD murF murE
7	0.001427376	4	6	21.39250493	Mur ligase family, glutamate ligase domain	Ligase murD murF murE
8	0.001985843	4	7	18.3364328	Mur ligase, central	Ligase murD murF murE
9	0.001985843	4	7	18.3364328	Mur-like, catalytic domain superfamily	Ligase murD murF murE
10	0.001985843	4	7	18.3364328	Mur ligase middle domain	Ligase murD murF murE
11	4.87E-06	8	15	17.11400394	Chaperone, and Clp protease	dnaK grpE dnaJ hspR groEL2 groES clpP1 tig
12	0.018605244	3	6	16.0443787	Serine protease	MMAR_1370 cphB clpP1
13	0.026646888	3	7	13.7523246	Biosynthesis of siderophore group nonribosomal peptides	mbtB MMAR_3693 mbtC
14	0.000559163	7	19	11.82217378	Chaperone	dnaK grpE dnaJ groEL2 groES rimM tig
15	0.010183059	4	11	11.66863905	Cell shape	murD murX murF murE
16	0.01374453	4	12	10.69625247	Translation protein, beta-barrel domain superfamily	rimM MMAR_2900 cysN typA
17	0.01374453	4	12	10.69625247	Peptidoglycan synthesis	murD murX murF murE
18	0.018007424	4	13	9.873463814	Biosynthesis of siderophore group nonribosomal peptides, and MbtH-like domain	mbtB mbtF MMAR_3693 mbtC
19	0.018007424	4	13	9.873463814	NAD binding domain of 6-phosphogluconate dehydrogenase	fadB2 proC serA1 gpdA2
20	0.022659134	4	14	9.168216399	Cell wall biogenesis/degradation	murD murX murF murE
21	0.01374453	5	21	7.640180332	Mixed, incl. mce/mlad, and permease mlae	MMAR_0162 MMAR_0165 yrbE6B mce6D mce6E
22	0.005196909	6	26	7.405097861	Cell cycle	ftsX murD murX murF murE tig
23	0.005196909	6	26	7.405097861	Cell division	ftsX murD murX murF murE tig
24	0.007428424	6	28	6.876162299	Phosphopantetheine attachment site, and Biosynthesis of siderophore group nonrib	fadD30 pks6 mbtB mbtF MMAR_3693 mbtC
25	0.003084466	7	33	6.806706114	Polyketide synthase, phosphopantetheine-binding domain, and Biosynthesis of side	fadD30 pks6 MMAR_3272 mbtB mbtF MMAR_3693 mbtC
26	0.001337958	10	55	5.834319527	Polyketide synthase, phosphopantetheine-binding domain, and MbtH-like domain	fadD30 pks6 drrA ppsD MMAR_1876 MMAR_3272 mbtB mbtF MMAR_3693 mbtC
27	0.002433398	10	68	4.718934911	Mixed, incl. polyketide synthase, phosphopantetheine-binding domain, and mbth-li	fadD30 pks6 drrA ppsD MMAR_1876 MMAR_3272 mbtB mbtF MMAR_3693 mbtC
28	0.001985843	13	109	3.827099506	Mixed, incl. polyketide synthase, phosphopantetheine-binding domain, and cytochr	fadD30 pks6 drrA ppsD MMAR_1876 MMAR_3024 pks12 mmpS4 MMAR_3272 mbtB mbtF MMAR_3693 mbtC
29	0.001337958	17	158	3.452587821	Mixed, incl. translation, and protein export	dnaK grpE dnaJ hspR groEL2 rpsH rpsE groES ftsX rimM trmD MMAR_2122 MMAR_2900 clpP1 tig MMAR_4096 pth
30	0.001337958	19	192	3.175449951	Mixed, incl. gene expression, and protein export	dnaK grpE dnaJ hspR groEL2 rpsH rpsE groES ftsX rimM trmD MMAR_1924 MMAR_2122 MMAR_2900 clpP1 tig
31	0.003084466	20	243	2.641049991	Cytoplasm	grpE dnaJ groEL2 proC groES gpdA2 rimM trmD pyrH argH ppnK murD murF murE glyS clpP1 tig pth purS gid

Table 5.10 GO pathway analysis performed of common genes significantly up in KO v WT (UKvWT) and down in COM v KO (DCvKO).
GO pathway analysis performed of common genes significantly upregulated in KOvWT and downregulated in COMvKO having log2FoldChange equal to or more than 1 and padj values less than 0.01. The table shows top 30 pathways according to the fold enrichment. Enrichment FDR cut off for this anlaysis was 0.05. Total genes involved in the pathway are given in column 3 and the number of analysed genes associated with that pathway are given in column 2. Pathways are ordered according to the fold enrichment score. Associated pathways are highlighted in one colour.

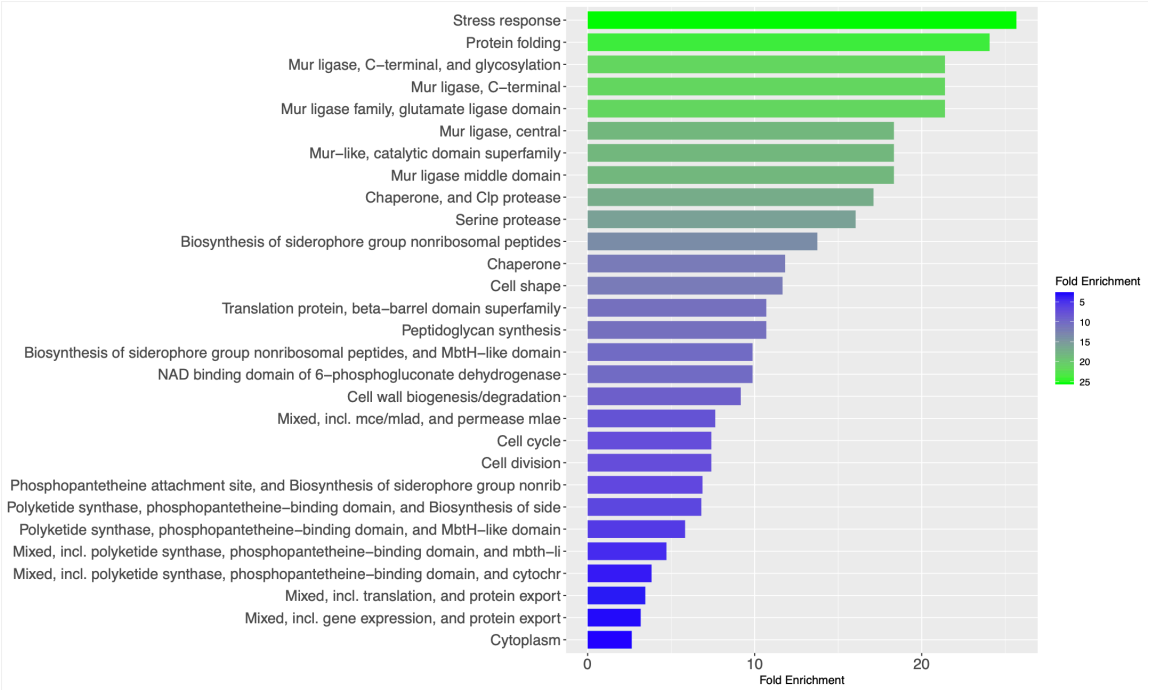


Figure 5.36 Bar plot showing the top pathways associated with the common genes up in KO v WT (UKvWT) and down in COM v KO (DCvKO).
Bar plot showing the top pathways associated with the common genes Up in KOvWT – down in COM v KO (Log2FC cut off 1, padj <0.01) and fold enrichment of each pathway. Analysis done using http://bioinformatics.sdstate.edu/.

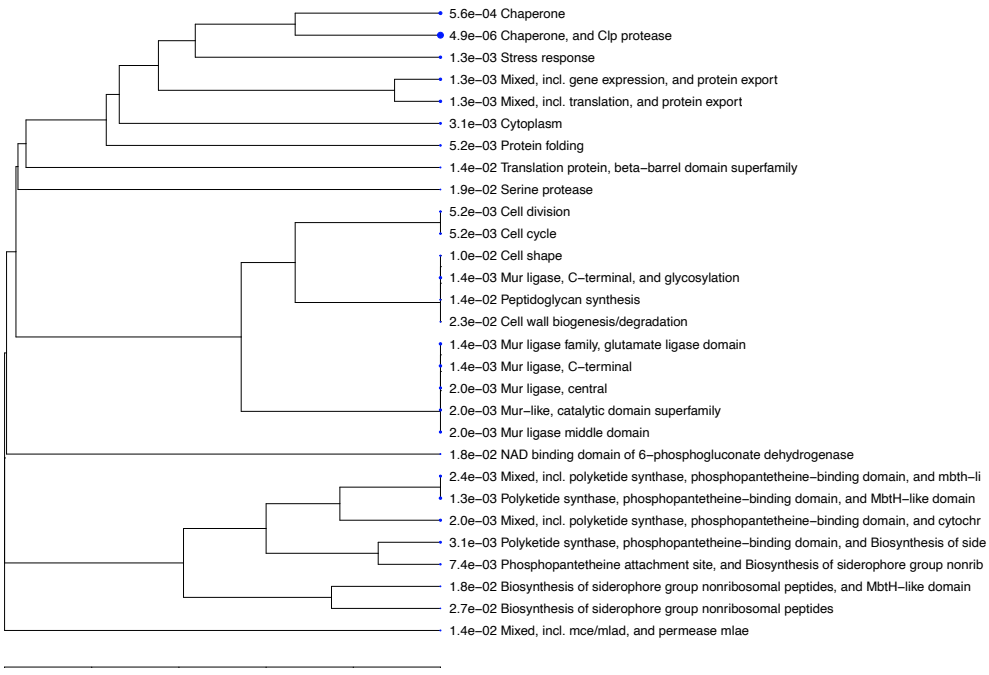


Figure 5.37 hierarchical clustering tree.
The hierarchical clustering tree showing the correlation of different pathways shown in figure 5.36. Pathways clustered together indicate the shared genes.

5.3.5.1.2 Analysis using *M. marinum* M homologous genes – log2FC between 1 – 0.585, padj value < 0.01

GO pathway analysis was also performed for the list of common genes UkvWT_DCvKO with log2FC between 1 and 0.585, padj <0.01. Out of 314 genes, 309 were found to have homologs in *M. marinum* M strain. The analysis was performed for 170 M homologous genes (used in section 5.3.5.1.1) and 309 M homologous genes. So, all the genes having log2FoldChange > 0.585 were included in this analysis.

Out of the top 60 pathways, the three pathways discussed in section A were still at the top of the list according to the fold enrichment (Figure 5.38). In addition to the genes identified in section 5.3.5.1.1, more genes were found to be associated with peptidoglycan pathway such as *murB, ddlA, uppP, ftsE, whiA, and mrsA/glmM*. The log2FC values and padj values of all the genes are given in Table 5.11. Similarly, more genes were identified to be associated with polyketide synthase and correlated pathways such as fadD22 upregulated in KOvWT with log2FC of 0.9589 and padj value 8.61e-07 and downregulated in COMvKO with log2FC of -0.9208 and padj value of 2.16e-06. Other genes observed to be involved in these pathways were *psk1, psk7, psk9, ppsE, rbfA, drrC*, and MMAR_0102 (Table 5.12). The full list of top 60 pathways obtained by GO pathway analysis and the genes associated with each pathway is given in supplementary material.

Apart from the three associated pathways discussed above, many genes were found to be involved in a variety of metabolic processes (highlighted in red in Figure 5.39). Many significantly upregulated genes in KOvWT were associated with purine nucleotide, ribonucleotide, and ribonucleotide triphosphate binding. Purine nucleotides are essential for cell survival because of their role in regulation of various metabolic process, energy transfer, and they play important role in DNA and RNA synthesis (highlighted in blue in Figure 5.39).

As discussed in sections 5.3.5.1.1 and 5.3.5.1.2 many genes involved in peptidoglycan synthesis, cell wall biogenesis, and polyketide pathways were identified in the GO pathway analysis. In the next section the role of these genes in their respective pathways is discussed further.

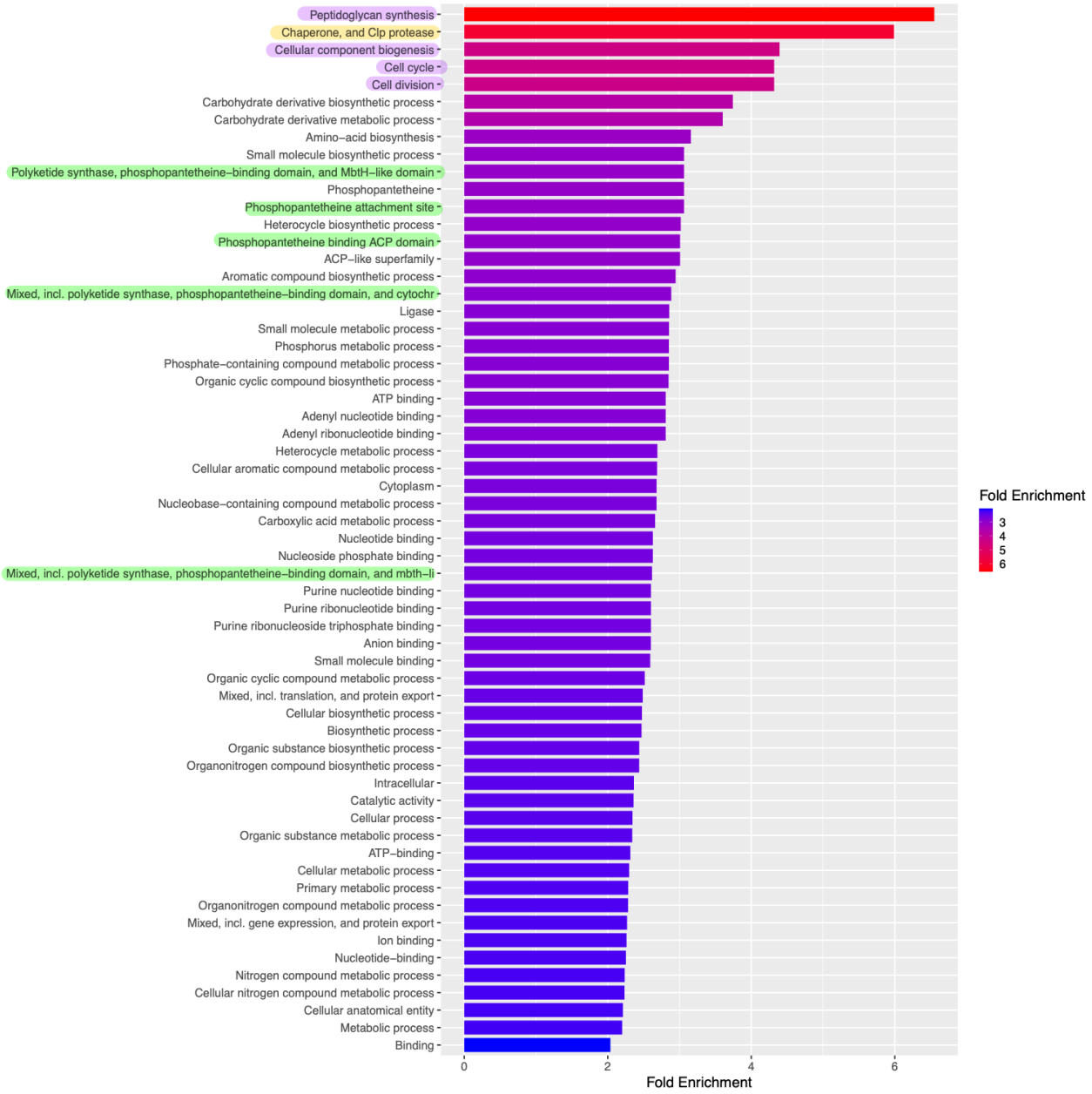


Figure 5.38 Bar plot showing the top pathways.
Bar plot showing the top pathways associated with the common genes up in KOvWT – down in COMvKO (Log2FC cut off 0.585, padj <0.01) and fold enrichment of each pathway. Analysis was done using http://bioinformatics.sdstate.edu/.

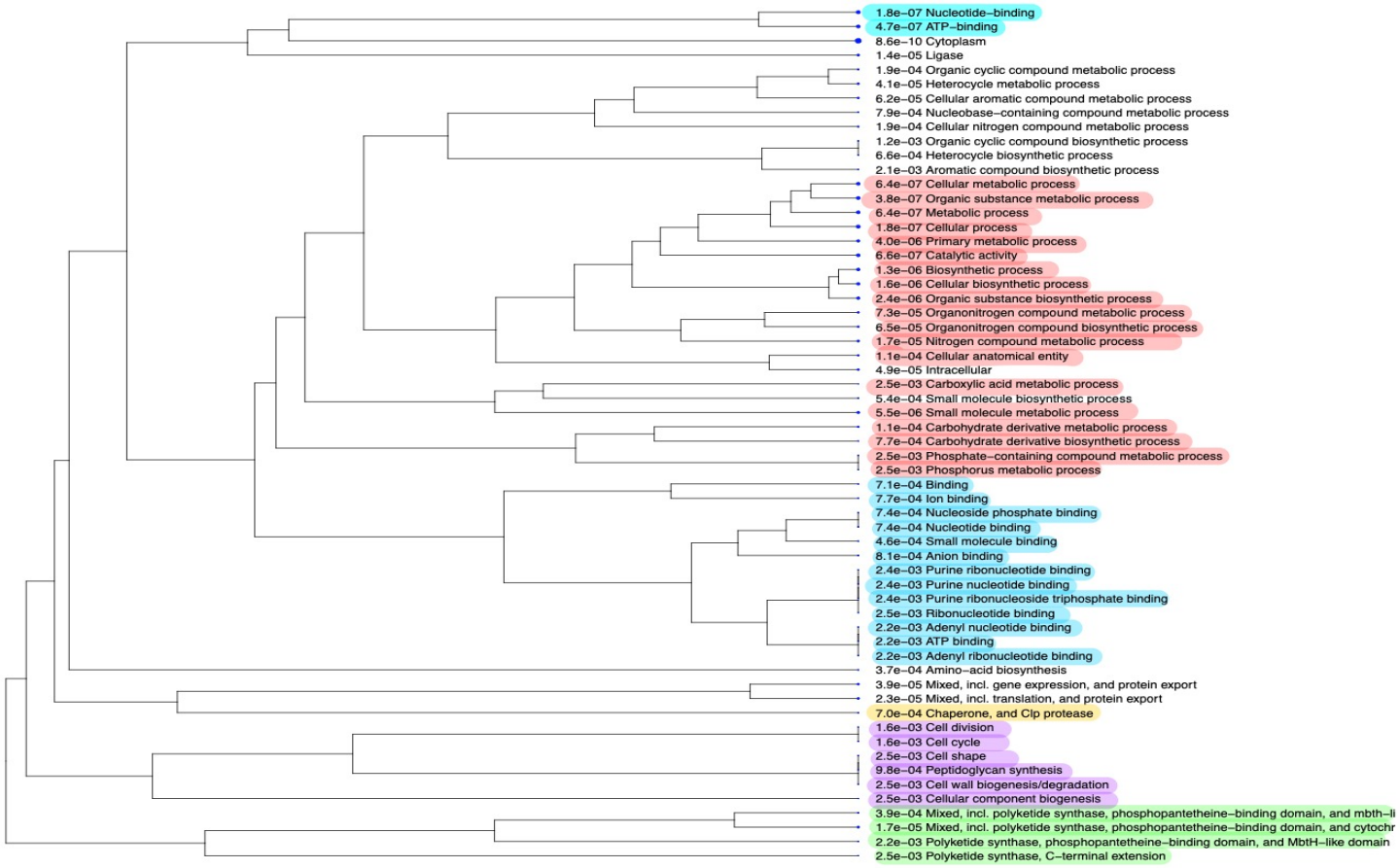


Figure 5.39 Hierarchical clustering tree.
The hierarchical clustering tree showing the correlation of different pathways shown in figure 5.38. Pathways clustered together indicate the shared genes.

Sr No.	ID	gene_name	description	log2FC_KOvWT	padj_KOvWT	log2FC_COMvKO	padj_COMvKO	M strain hoomolog	M gene_name	Rv homolog	Rv gene_name
2) Cell wall, cell division, and peptidoglycan synthesis pathway											
1	MM1218R_RS16000	MM1218R_RS16000	UDP-N-acetylmuramoyl-L-alanine--D-glutamate ligase	1.119452584	1.10E-07	-1.373295588	4.37E-11	MMAR_3195	murD	Rv2155c	murD
2	MM1218R_RS16015	MM1218R_RS16015	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2%2C6-diaminopimelate ligase	1.240438471	4.99E-14	-1.168742518	1.05E-12	MMAR_3198	murE	Rv2158c	murE
3	MM1218R_RS16010	murF	UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase	1.116920697	5.60E-16	-1.033939205	2.99E-14	MMAR_3197	murF	Rv2157c	murF
4	MM1218R_RS16005	MM1218R_RS16005	phospho-N-acetylmuramoyl-pentapeptide-transferase	1.150600976	7.07E-16	-1.257558657	6.40E-19	MMAR_3196	murX	Rv2156c	murX / mraY
5	MM1218R_RS07215	ftsX	permease-like cell division protein FtsX	1.444088426	3.31E-12	-1.873711218	9.34E-20	MMAR_1531	ftsX	Rv3101c	ftsX
6	MM1218R_RS03815	MM1218R_RS03815	UDP-N-acetylmuramate dehydrogenase	0.59144437	0.003792764	-0.781739811	8.88E-05	MMAR_0808	murB	Rv0482	murB
7	MM1218R_RS15730	MM1218R_RS15730	undecaprenyl-diphosphate phosphatase	0.7572032	1.99E-05	-0.889053423	3.81E-07	MMAR_3116	uppP	Rv2136c	Rv2136c
8	MM1218R_RS08665	MM1218R_RS08665	D-alanine--D-alanine ligase	0.71511553	3.65E-05	-0.585318264	0.000759145	MMAR_1733	ddlA	Rv2981c	ddlA
9	MM1218R_RS05355	MM1218R_RS05355	phosphoglucosamine mutase	0.700991415	7.76E-08	-0.917839615	4.60E-14	MMAR_1108	mrsA	Rv3441c	mrsA / GlmM
10	MM1218R_RS07210	ftsE	cell division ATP-binding protein FtsE	0.968545599	1.32E-10	-0.96309736	1.19E-10	MMAR_1530	ftsE	Rv3102c	ftsE
11	MM1218R_RS11130	whiA	DNA-binding protein WhiA	0.918271714	4.84E-09	-0.916190977	4.23E-09	MMAR_2525	whiA	Rv1423	whiA
12	MM1218R_RS12670	scpB	SMC-Scp complex subunit ScpB	0.76991123	1.41E-06	-0.83287638	1.37E-07	MMAR_2525	MMAR_2525	Rv1710	scpB
13	MM1218R_RS10830	ruvX	Holliday junction resolvase RuvX	0.822337556	3.62E-06	-0.620476566	0.000502615	MMAR_2171	MMAR_2171	Rv2554c	Rv2554c

Table 5.11 Detail of proteins associated with cell wall, cell division, and peptidoglycan pathways.
The log2FC and padj values of KOvWT and COMvKO comparison are given. The homologous genes of M strain and H37Rv was determined by Roary analysis. The gene description was taken from NCBI database.

Sr No.	ID	gene_name	description	log2FC_KOvWT	padj_KOvWT	log2FC_COMvKO	padj_COMvKO	M strain hoomolog	M gene_name	Rv homolog	Hv gene_name
3) Polyketide synthase, phosphopantetheine-binding domain and related pathways											
1	MM1218R_RS15275	MM1218R_RS15275	type I polyketide synthase	1.302046478	2.18E-13	-1.143302247	1.26E-10	MMAR_3025	pks12	Rv2048c	pks12
2	MM1218R_RS03315	MM1218R_RS03315	acyltransferase domain-containing protein	1.077908348	8.93E-10	-1.389236115	1.40E-15	MMAR_0707	pks6	Rv0405	pks6
3	MM1218R_RS03310	MM1218R_RS03310	AMP-binding protein	1.095162496	1.39E-09	-1.234362729	5.42E-12	MMAR_0706	fadD30	Rv0404	fadD30
4	MM1218R_RS18800	MM1218R_RS18800	non-ribosomal peptide synthetase	1.252148993	2.02E-16	-1.008048796	8.46E-12	MMAR_3689	mbtB	Rv2383c	mbtB
5	MM1218R_RS18825	MM1218R_RS18825	polyketide synthase	1.144505199	2.07E-06	-1.141843886	1.04E-06	MMAR_3694	mbtC	Rv2382c	mbtC
6	MM1218R_RS18805	MM1218R_RS18805	non-ribosomal peptide synthetase	1.02818625	2.17E-14	-1.156275894	3.81E-18	MMAR_3690	mbtF	Rv2379c	mbtF
7	MM1218R_RS08875	MM1218R_RS08875	ATP-binding cassette domain-containing pro	1.357817607	6.93E-20	-1.251546643	2.58E-17	MMAR_1771	drrA	Rv2936	drrA
8	MM1218R_RS08885	MM1218R_RS08885	type I polyketide synthase	1.104084205	4.39E-21	-1.049844967	2.65E-19	MMAR_1773	ppsD	Rv2934	ppsD
9	MM1218R_RS15270	MM1218R_RS15270	sugar epimerase family protein	1.023846923	2.23E-15	-1.021015564	1.88E-15	MMAR_3024	MMAR_3024	Rv2047c	Rv2047c
10	MM1218R_RS16580	MM1218R_RS16580	fatty acyl-AMP ligase	1.169642525	1.39E-12	-1.151996313	1.27E-12	MMAR_3272	MMAR_3272	Homolog not found	
11	MM1218R_RS09425	MM1218R_RS09425	AMP-binding protein	1.034251926	3.21E-07	-1.166140328	3.91E-09	MMAR_1876	MMAR_1876	Homolog not found	
12	MM1218R_RS18820	MM1218R_RS18820	alpha/beta fold hydrolase	1.055123339	0.000493004	-1.129106288	0.000113483	MMAR_3693	MMAR_3693	Homolog not found	
13	MM1218R_RS15905	MM1218R_RS15905	MmpS family transport accessory protein	1.711255365	0.0066401	-1.938133552	0.001714755	MMAR_3175	mmpS4	Homolog not found	
14	MM1218R_RS11680	MM1218R_RS11680	SDR family NAD(P)-dependent oxidoreducta	1.333811156	1.30E-34	-1.291623225	1.59E-32	MMAR_2340	pks5	Rv3825c	pks2
15	MM1218R_RS09050	rimM	ribosome maturation factor RimM	1.822126957	4.10E-21	-1.659716708	5.14E-18	MMAR_1801	rimM	Rv2907c	rimM
16	MM1218R_RS08810	MM1218R_RS08810	p-hydroxybenzoic acid--AMP ligase FadD22	0.958905196	8.61E-07	-0.920884516	2.16E-06	MMAR_1761	fadD22	MMAR_1761	fadD22
17	MM1218R_RS08815	MM1218R_RS08815	type I polyketide synthase	0.825312233	2.98E-07	-0.67288797	3.23E-05	MMAR_1762	pks15/1	Rv2946c	pks1
18	MM1218R_RS12365	MM1218R_RS12365	SDR family NAD(P)-dependent oxidoreducta	0.791600478	6.90E-08	-0.729461882	6.23E-07	MMAR_2471	pks7	Rv1661	pks7
19	MM1218R_RS12375	MM1218R_RS12375	type I polyketide synthase	0.965376967	2.60E-07	-0.88289782	2.11E-06	MMAR_2473	pks9	Rv1664	pks9
20	MM1218R_RS08880	MM1218R_RS08880	type I polyketide synthase	0.781808274	5.07E-07	-0.595277359	0.000149825	MMAR_1772	ppsE	Rv2935	ppsE
21	MM1218R_RS09525	rbfA	30S ribosome-binding factor RbfA	0.824243758	3.05E-11	-0.832577013	1.26E-11	MMAR_1895	rbfA	Rv2838c	rbfA
22	MM1218R_RS08865	MM1218R_RS08865	ABC transporter permease	0.82657019	2.36E-08	-0.724477905	8.95E-07	MMAR_1769	drrC	Rv2938	drrC
23	MM1218R_RS00470	MM1218R_RS00470	type I polyketide synthase	0.649977249	4.81E-04	-0.703447848	0.000110029	MMAR_0102	MMAR_0102	Homolog not found	

Table 5.12 Detail of proteins associated with polyketide synthase, phosphopantetheine-binding, and related pathways.
The log2FC and padj values of KOvWT and COMvKO comparison are given. The homologous genes of M strain and H37Rv was determined by Roary analysis. The gene description was taken from NCBI database.

5.3.5.1.3 Role of the genes associated with cell wall biogenesis identified in pathway analysis in their respective pathways.

The cell wall of *Mycobacterium tuberculosis* has a distinctive structure and composition. It is composed of an inner layer of peptidoglycan which is just above the plasma membrane, a covalently attached layer of arabinogalactan, a thick homogenous layer of unique mycolic acid and an outer heterogenous layer of capsule. The outer layer is composed of a variety of non-covalently attached lipids which includes lipoarabinomannan (LAM), lipoglycans, lipomannan, glycosylated or non-glycosylated lipoproteins, trehalose-6-monomycolate and trehalose-6-6'-dimycolate (Figure 5.40). The peptidoglycan layer of mycobacterial cell wall is essential for cell wall integrity and pathogenesis of the bacilli. Mur ligases are an important class of enzymes involved in the biosynthesis and recycling of peptidoglycan of mycobacterial cell wall. Studies have shown that the multiple *mur* ligases could be a promising target for non-antibiotic treatment of tuberculosis (Maitra et al., 2019).

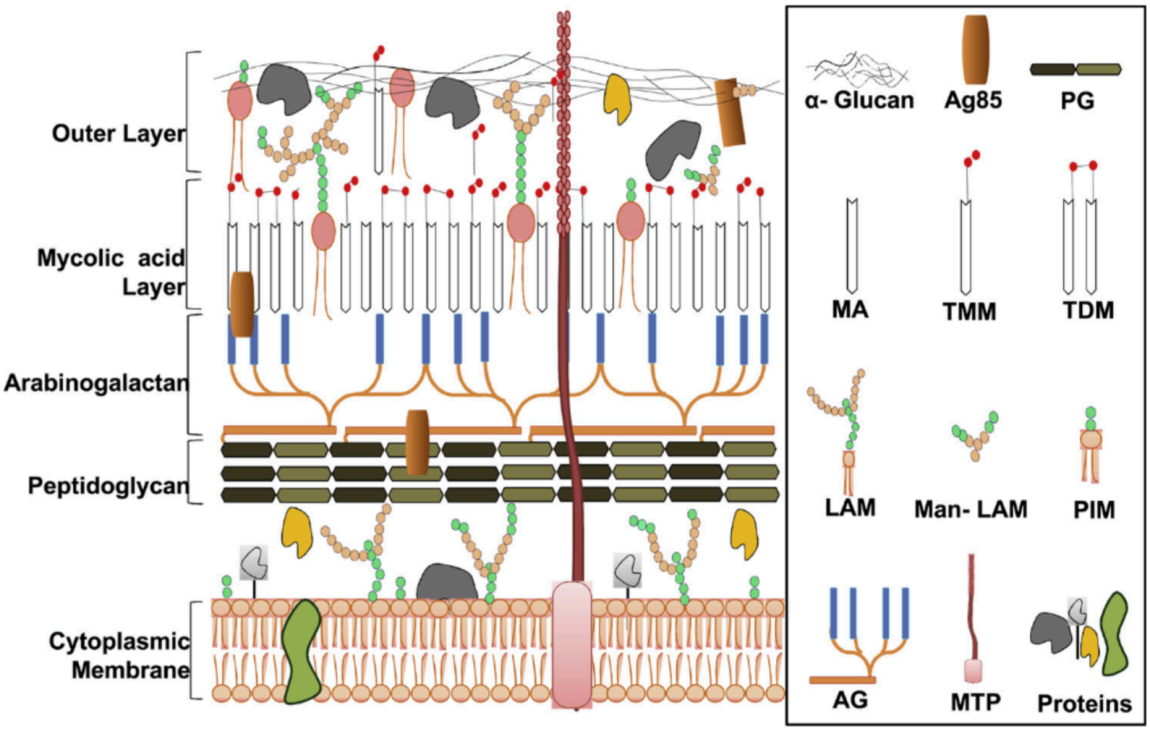


Figure 5.40 Diagrammatic representation of cell envelope of *Mycobacterium tuberculosis*.

PG (peptidoglycan), MA (mycolic acid), TMM (tre-halose-6-monomycolate), TDM (trehalose 6, 6'-dimycolate), LAM (lipoarabinomannan), Man-LAM (mannose-capped lipoarabinomannan), PIMs (phosphatidylinositol manno-sides). Image taken from (Vinod et al., 2020).

1) Peptidoglycan biogenesis

The full schematic pathway of biogenesis, degradation and recycling of mycobacterial peptidoglycan is shown in Figure 5.41. The synthesis begins with fructose-6-phosphate conversion to UDP-N-acetylglucosamine. Out of the three enzymes involved in this step, one enzyme GlmM (Rv3441c) was found to be upregulated in KOvWT by log2FC of 0.7 and padj value 7.76e-08 (Table 5.11, row 9).

In second step UDP-N-acetylglucosamine is converted to (UDP-MurNAc) UDP-N-acetylmuramate. This step is catalyzed by *murA* and *murB*. *murB* is a type of NADPH dependent oxidoreductase. This enzyme was found to be upregulated by a log2FC of 0.59 and padj value of 0.003 (Table 5.11, row 6).

The next step is the assembly of the peptidoglycan monomer unit which is catalyzed by four essential ATP-dependant *mur* ligases namely *murC, murD, murE, and murF*. Out of these four essential ligases, three were found to be significantly differentially expressed in KOvWT. *murD, murE and murF* upregulated in KOvWT by a log2FC values > 1 and padj value < 1e-07 (Table 5.11, rows 1-3).

The hydrophilic peptidoglycan precursors are transported from cytoplasm across hydrophobic plasma membrane. This step is catalyzed by MraY transferase. MraY is the Mtb homologous of MurX (MM1218R_RS16005) which is upregulated in KOvWT by a log2FC of 1.15.

The expression of all the six genes discussed above was found to be reversed in COMvKO by nearly the equivalent log2FC which suggests that the absence of Cpn60.1 directly or indirectly caused the upregulation of these essential enzymes critical for the peptidoglycan synthesis. This suggests that Cpn60.1 might have an essential direct or indirect effect on the cell wall or peptidoglycan integrity and that in the absence of Cpn60.1, degradation / recycling of peptidoglycan is increased which might lead to the overexpression of key enzymes involved in the synthesis of peptidoglycan.

2) Polyketide synthases

The mycobacterial genome encodes a large number of polyketides which play a fundamental role in the biosynthesis of complex heterogenous lipids of cell envelope.

There are three types of mycobacterial polyketide synthases. Each type is involved in different steps of biosynthesis of cell wall lipids. The cell envelope glycolipids play an essential role in strengthening the cell envelope and provide resistance against several antibiotics. Studies have shown that for the biosynthesis of phenolic glycolipids (PGLs) and DIMs (Phthiocerol dimycocerosate), an intermediate compound PHPA (p-hydroxyphenylalkanoate) is synthesized by the combined catalytic activity of FadD22 and Pks15/1 (Siméone et al., 2010). The pathway is shown in Figure 5.42.

The genes encoding both the enzymes FadD22 and PKS 15/1 were shown to be upregulated in KOvWT by nearly same order of mangnitude (Table 5.12). Several genes such as *pks5, ppsD, ppsE, mbtB, mbtC, mbtF*, and *mmpS4* upregulated in KOvWT were found to be involved in the transport of complex lipids across the plasma membrane. Various other *pks* and associated genes were also significantly upregulated e.g., *psk12, pks6, pks7*, *pks9, drrA,* and *drrC* (Table 5.12). Overall, the results show that the absence of Cpn60.1 causes significant upregulation of various essential genes involved in the synthesis and export of essential cell wall lipids which indicates the importance of Cpn60.1 in the maintenance of mycobacterial cell wall.

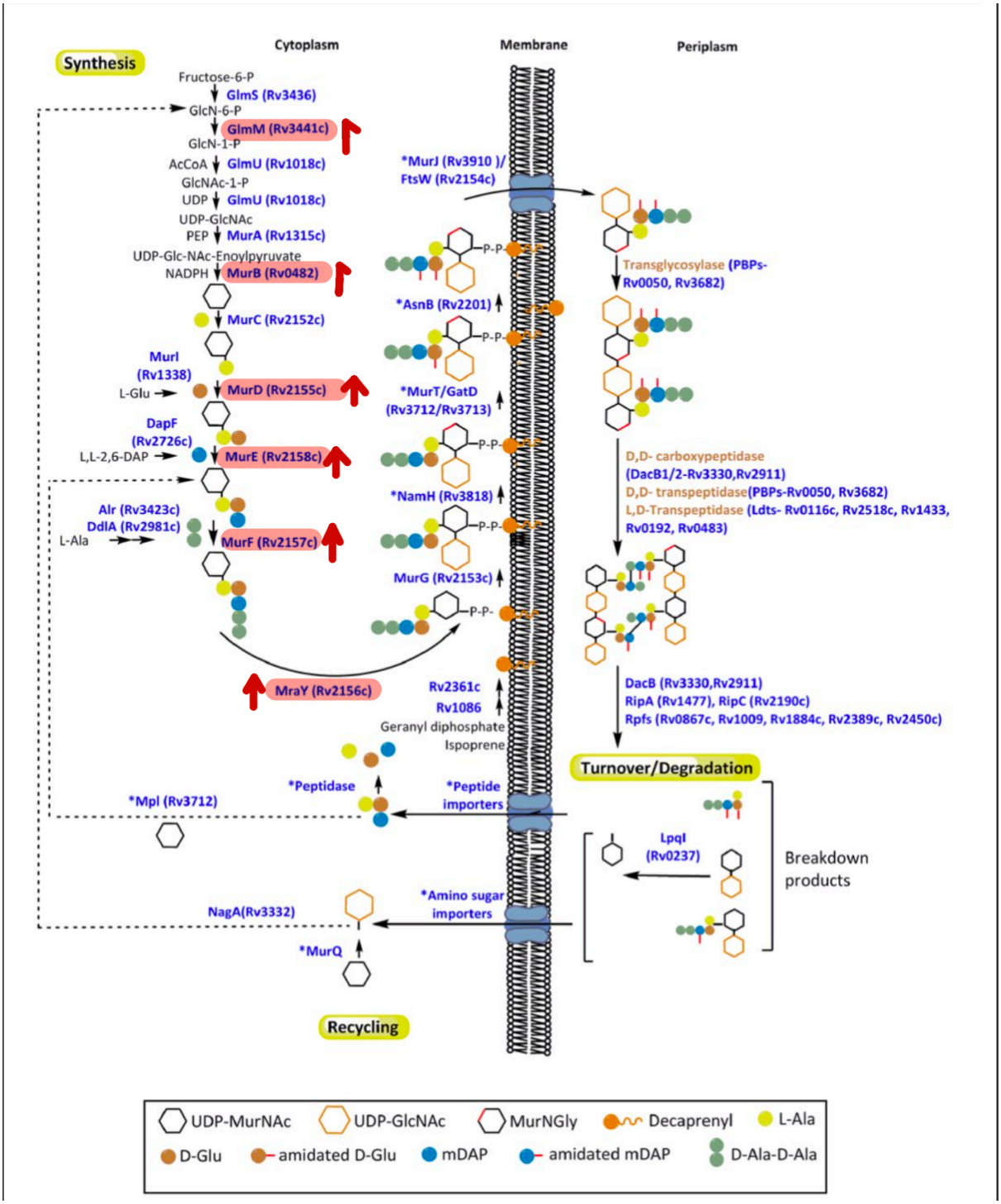


Figure 5.41 Peptidoglycan biosynthesis, degradation, and recycling pathways.
An overview of peptidoglycan biosynthesis, degradation, and recycling pathways representing information known for mycobacterial species. Image adapted from (Maitra et al., 2019). The genes UKvWT_DCvKO log2FC >=1, padj <0.01 are highlighted in red and marked with full arrowhead. The genes UKvWT_DCvKO log2FC between 1-0.05 padj <0.01 are highlighted in red and marked with half arrowhead.

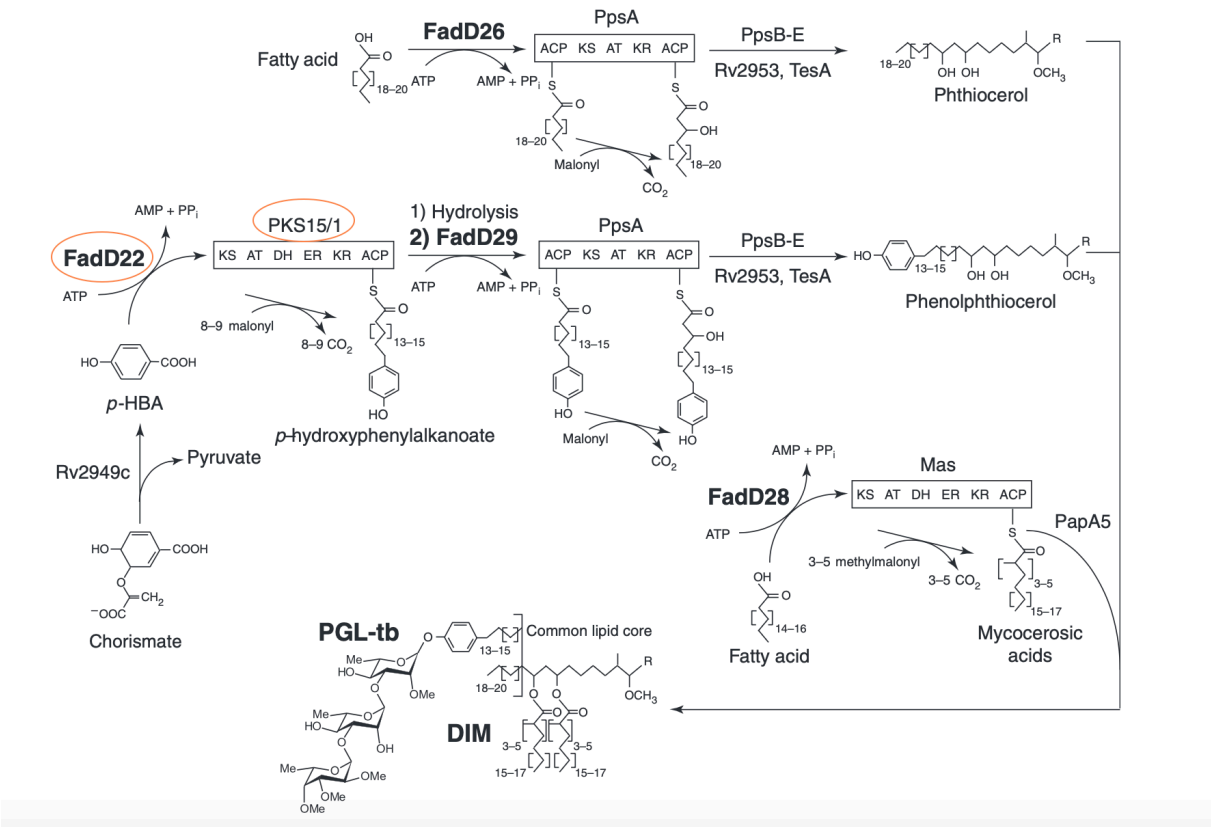


Figure 5.42 Schematic sketch of the role of FadD enzymes in the biosynthesis of cell envelope PGLs and DIMs.
FadD22 and PKS15/1 work together in this pathway. Figure adapted from (Siméone et al., 2010).

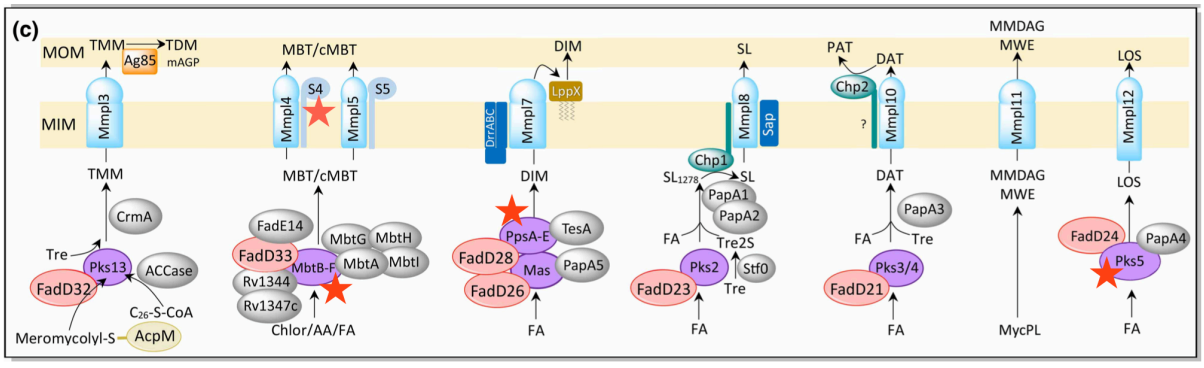


Figure 5.43 Schematic diagram showing the transport mechanism of complex lipids across the cell membrane in *M. tuberculosis*.
Stars indicate the enzymes found to be upregulated in KOvWT i.e., Pks5, PpsD, PpsE, MbtB, MbtC, MbtF, and MmpS4. FA (fatty acids) and Tre (trehalose). Image taken from (Batt et al., 2020).

5.3.5.2 Pathway analysis of common genes DKvWT_UCvKO

1) Analysis using *M. marinum* M homologous genes

The 315 common genes Down in KOvWT but Up in COMvKO having log2FoldChange values cut off -1 and padj value 0.05 were used in the analysis. From this list 11 genes having padj values greater than 0.01 were removed. The remaining 304 genes found to have 270 homologous genes in *M. marinum* M. The anlaysis was performed using those 270 genes.

Among the highest fold enrichment pathways, nitrotoluene degradation, xylose isomerase like TIM barrel pathway, and various enzyme were included. Pathways involving carbon monoxide dehydrogenase, acetoacetate decarboxylase related pathways were also in the top 10 according to fold enrichment (Figure 5.44). Some virulence factor genes were also downregulated in KOvWT. Figure 5.42 illustrates the cluster diagram of relatedness of different pathways. Top 10 pathways and the associated genes are shown in Table 5.13. The full list of pathways and associated genes is given in supplementary data.

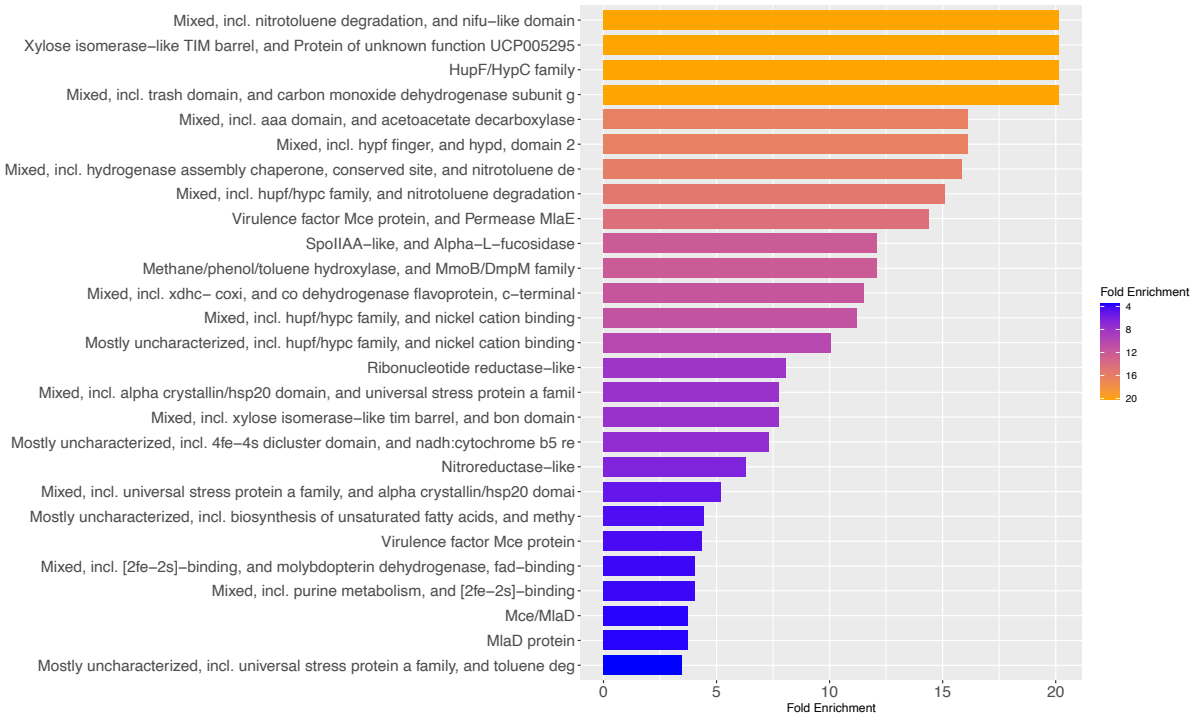


Figure 5.44 Bar plot showing the top pathways associated with the common genes DKvWT – UCvKO.
Bar plot showing the top pathways associated with the common genes down in KOvWT – up in COMvKO (Log2FC cut off -1, padj <0.01) and fold enrichment of each pathway. Analysis was done using http://bioinformatics.sdstate.edu/.

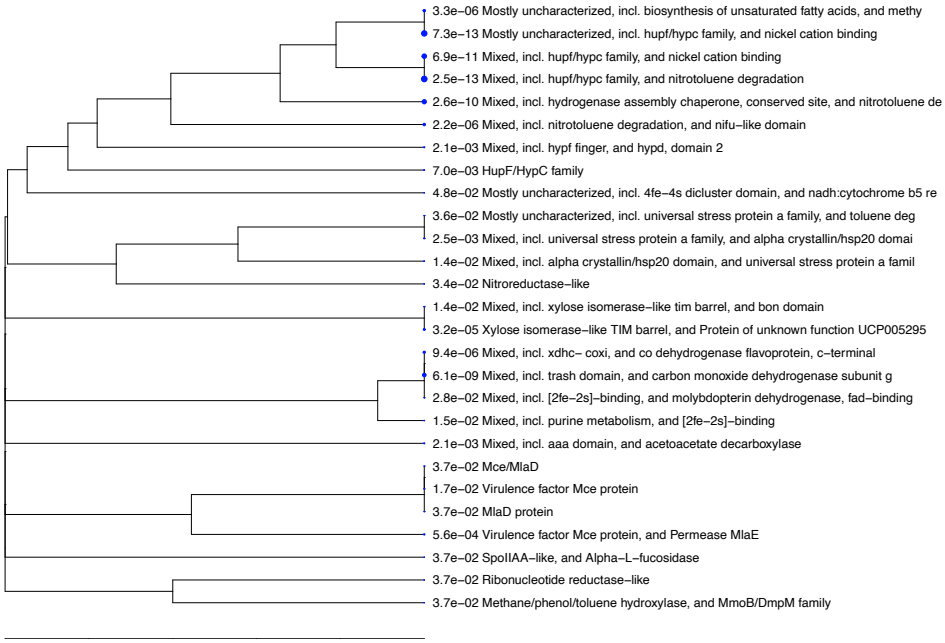


Figure 5.45 The hierarchical clustering tree showing the correlation of different pathways shown in figure 5.44..

	Enrichment FDR	nGenes	Pathway Genes	Fold Enrichmer	Pathway	Genes
2	2.22E-06	6	6	20.1598513	Mixed, incl. nitrotoluene degradation, and nifu-like domain	MMAR_1865 MMAR_1866 MMAR_1867 MMAR_1868 hyaB hyaA
3	3.19E-05	5	5	20.1598513	Xylose isomerase-like TIM barrel, and Protein of unknown function UCP005295	MMAR_2509 MMAR_2510 MMAR_2511 MMAR_2512 MMAR_2513
4	0.007004027	3	3	20.1598513	HupF/HypC family	hypC MMAR_1861 MMAR_1863
5	6.09E-09	8	8	20.1598513	Mixed, incl. trash domain, and carbon monoxide dehydrogenase subunit g	coxM coxS coxL MMAR_0658 MMAR_0659 MMAR_0660 MMAR_0661 MMAR_0662
6	0.002091852	4	5	16.12788104	Mixed, incl. aaa domain, and acetoacetate decarboxylase	MMAR_4818 MMAR_4819 MMAR_4820 MMAR_4821
7	0.002091852	4	5	16.12788104	Mixed, incl. hypf finger, and hypd, domain 2	hypD hypF MMAR_1863 hyaD
8	2.58E-10	11	14	15.83988317	Mixed, incl. hydrogenase assembly chaperone, conserved site, and nitrotoluene de	hypD hypC hypF MMAR_1863 hyaD MMAR_1865 MMAR_1866 MMAR_1867 MMAR_1868 hyaB hyaA
9	2.53E-13	15	20	15.11988848	Mixed, incl. hupf/hypc family, and nitrotoluene degradation	MMAR_1856 hypD hypC MMAR_1860 MMAR_1861 hypF MMAR_1863 hyaD MMAR_1865 MMAR_1866 MMAR_1867 MMAR_1868 hyaB hyaA MMAR_1874
10	0.000561243	5	7	14.39989379	Virulence factor Mce protein, and Permease MlaE	yrbE3A_1 mce3A_1 mce3B_1 lprM_1 mce3F_1
11	0.037415621	3	5	12.09591078	SpoIIAA-like, and Alpha-L-fucosidase	MMAR_2577 MMAR_2847 MMAR_2848

Table 5.13 top 10 pathways according to the fold enrichment.
GO pathway analysis performed of common genes significantly downregulated in KOvWT and upregulated in COMvKO having log2FoldChange equal to or more than -1 and padj values less than 0.01. The table shows top 10 pathways according to the fold enrichment.

5.3.6 PE/PPE family genes, Hypothetical proteins, and DUF domain containing proteins

Several genes encoding PE/PPE family proteins, DUF domain family proteins and hypothetical proteins were also significantly upregulated or downregulated in KOvWT. Several genes showed reverse in the expression magnitude in the complemented strain. Some up and down regulated DUF family protein encoding genes and PE/PPE family proteins encoding genes were highlighted in Figure 5.27. We don't know the exact function of these genes, neither any information about such genes was found by the GO pathway analysis. However, this shows that the absence Cpn60.1 in the cell shifts the transcriptome profile of whole cell affecting the expression of hundreds of genes directly and or indirectly.

5.3.7 Relative expression of chaperonin genes in WT samples

As discussed in chapter 4 (section 4.1.1 & 4.1.2) that the two bicistronic *cpn10–cpn60.1* genes of *M. tuberculosis* were reported to express differentially both under normal and stressed conditions (Fischer et al., 1993; Hu et al., 2008; Sato et al., 2008). Experiments with *M. tuberculosis* showed that the *cpn10* promoter had the strongest promoter activity followed by *cpn60.2* whereas no promoter activity was found upstream to cpn60.1 (Aravindhan et al., 2009; Hu et al., 2008). In *M. smegmatis* under normal growth conditions, *cpn60.2* was found to be the most highly expressed gene, followed by the co-chaperonin *cpn10*, and then *cpn60.1* (Rao & Lund, 2010).

In *M. marinum,* the fragment upstream to *cpn60.2* showed the strongest promoter activity followed by that of *cpn10* whereas no promoter activity was observed in the intergenic region of *cpn60.1* (chapter 4, section 4.4.3). This suggests that *cpn60.2* might be expressed at the highest levels in *M. marinum* followed by *cpn10*. To investigate this hypothesis, the read counts of WT samples of RNA-seq data were normalized for TPM values (transcripts per million) so that the intra sample comparison of genes expression could be done. The bar graph was plotted using the mean TPM values of the triplicate samples of WT 1218R. Mm*cpn60.2* was found to be expressed at levels higher than the expression of other chaperonin genes followed by *cpn10* and *cpn60.1* (consistent with the observations of Rao & Lund, 2010. The bicistronic *cpn10-cpn60.1* was observed to be expressed differentially; the expression levels of *cpn10* were nearly

doubled the levels of *cpn60.1* (Figure 5.46). The expression of *cpn60.2* was slightly more than 2-fold higher than *cpn10*. To compare the RNA-seq results of *M. marinum* with that of *M. tuberculosis*, the data from the experiments of Shell et al., 2015 was used to plot the bar graph (Figure 5.47). The RPKM values of the triplicate samples of WT H37Rv showed that *cpn10* was expressed at highest levels followed by *cpn60.2* and *cpn60.1* (consistent with the results of Aravindhan et al., 2009; Hu et al., 2008.

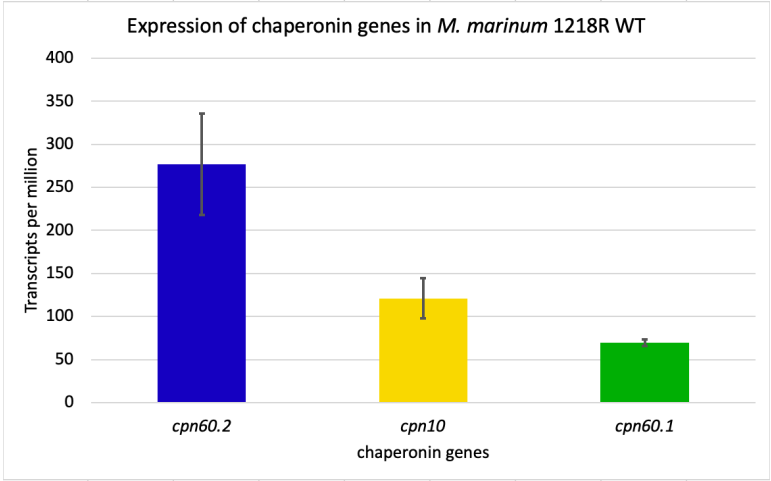


Figure 5.46 Relative expression of chaperonin genes in WT samples of *M. marinum* 1218R. Y-axis shows the mean TPM values of the triplicate WT samples of *M. marinum*. Expression of *cpn60.2* was found to be the highest of all the chaperonin genes followed by *cpn10* and *cpn60.1*. *cpn10* and *cpn60.1* showed differential expression. Error bars show standard deviation.

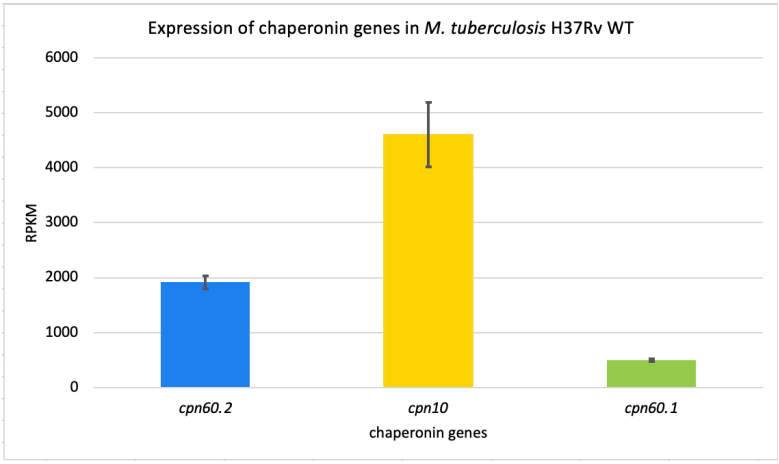


Figure 5.47 Relative expression of chaperonin genes in WT samples of *M. tuberculosis* H37Rv. Y-axis shows the mean RPKM values of the triplicate WT samples of *M. tuberculosis*. Expression of *cpn10* was found to be the highest of all the chaperonin genes followed by *cpn60.2* and *cpn60.1*. *cpn10* and *cpn60.1* showed differential expression. Error bars show standard deviation. The data is taken from the experiments of (Shell et al., 2015a).

5.4 Discussion

The RNA-seq analysis data was high quality which provided important insights into the role of MmCpn60.1 in the cell. Overall, 184 genes were found to be significantly upregulated in KO but downregulated in complement strain with log2FoldChange higher than 1. The padj values of all these significantly differentially expressed common genes (up in KO but down in COM) were less than 0.05. Although majority of them had extremely low padj values i.e., less than 0.0001. The gene ontology pathway analysis of these genes showed that most of the genes were involved in key mechanisms of the cell such as stress response, cell wall and lipid biogenesis.

The deletion of Mm*cpn60.1* from the chromosome strongly affects the overall stress response of the cells which could either be because of the ability of Cpn60.1 to function as a chaperone in its native environment and that its deletion caused an upregulation of other stress response proteins to compensate the activity of Cpn60.1. It could also be due to an indirect effect of absence of Cpn60.1. Cpn60.1 might be involved in some specific activities which in the absence of Cpn60.1 cause the increase in the stress level of the cells. We cannot infer from the results of RNA-seq data either the effect was direct or indirect and which hypothesis is true. RNAseq data cannot give any information about the other possible proposed roles for Cpn60.1 in immune signalling as discussed in chapter 1 (section 1.4.6). To investigate this macrophage infections or infection assays of the host organism could be performed.

It was also discussed in detail that absence of Cpn60.1 causes an upshift of several enzymes which are crucial for the biosynthesis and transport of peptidoglycan. This indicates that Cpn60.1 might be involved in the maintenance of cell shape, cell division, and cell wall and that the absence of Cpn60.1 might cause an increase in the peptidoglycan degradation which in turn cause an upregulation of enzymes involved in the biogenesis of peptidoglycan. The significant upregulation of *mur* genes in KO strain and the downregulation of same genes in COM strain is interesting because most of them were found to play major role in the peptidoglycan biogenesis. Further investigation of the regulation of these genes would be helpful in understanding the role of Cpn60.1 in cell wall biogenesis.

A strong effect on the regulation of polyketide synthases and related pathways also indicates that Cpn60.1 might be associated with the lipid profile of mycobacteria. This corelates with the findings of (Ojha et al., 2005) where they demonstrated that *M. smegmatis cpn60.1* KO mutant was found to be defective in biofilm formation and that Cpn60.1 might be involved in a regulatory function in the synthesis of mycolic acid biogenesis. Two important enzymes FadD22 and PKS 15/1 which catalyse the initial steps of the biosynthesis of cell envelope PGLs and DIMs were significantly upregulated in KO and downregulated in COM strain. And the significant differential expression of several genes involved in the transport mechanism of complex lipids across cell membrane is interesting to investigate further to elucidate the role of Cpn60.1 in cell wall biogenesis maintenance.

To test if Cpn60.1 is involved in the lipid profile of the cells, experiments could be performed to look at the components of cell wall or lipid profiles of the WT, KO, and complemented strains. If the cell walls of Δ*cpn60.1* is found to be impaired, then it could cause a direct impact on mycobacterial pathogenesis because cell wall integrity is a crucial part of Mtb survival in the host.

TPM values of the WT samples showed that Mm*cpn60.2* has highest expression of all the chaperonin genes in *M. marinum*. This is consistent with the results of promoter probe study (Chapter4) where we observed that the Mm*cpn60.2* promoter had the stronger promoter activity than that of Mm*cpn10* both under normal and stressed conditions. Differential expression of Mm*cpn10* and Mm*cpn60.1* is consistent with the studies on other mycobacterial species reviewed in Chapter 4.

6 Expression and complementation analysis of *M. marinum* chaperonins in *E. coli*

6.1 Introduction

In mycobacteria, all the genetic analysis studies done so far have repeatedly demonstrated that out of the two chaperonin homologues *cpn60.2* is essential whereas *cpn60.1* is dispensable (Hu et al., 2008; Matern et al., 2020; Ojha et al., 2005; Rifat et al., 2021). As the chaperonin activity is essential for the survival of the cell, the indispensable nature of *cpn60.2* suggests that Cpn60.2 might perform the essential housekeeping chaperonin role in mycobacteria. This raises the question of the role of non-essential homologue *cpn60.1*. In the RNA-seq analysis (chapter 5), the levels of expression of *cpn60.2* were found to be increased in the Δ*cpn60.1* 1218R strain suggesting that Cpn60.1 might perform some chaperonin activity in *M. marinum* therefore upon deletion of *cpn60.1*, the expression of *cpn60.2* is increased to compensate the chaperonin requirements by the cell.

To get insight into the chaperonin function of mycobacterial chaperonins, previous studies investigated the ability of the mycobacterial chaperonins to complement the loss of GroEL in *E. coli*. Fan et al., 2012 studied the complementation ability of chaperonins of *M. tuberculosis* and *M. smegmatis* in *E. coli*. It was demonstrated that *cpn60.2* of both the species could complement the loss of GroEL in *E. coli* in either combination with *groES* of *E. coli* or *cpn10* of the respective species at normal growth temperature (37°C). In case of the combination of *M. tuberculosis cpn60.2* (Mtb*cpn60.2*) and *groES*, poorer complementation was observed under similar growth conditions. On the other hand, *cpn60.1* of both the species could not show complementation neither in combination with *groES* nor with *cpn10* (Fan et al., 2012). One reason for this could be the low expression levels of Cpn60.1 in *E. coli*. Several attempts were made to improve the expression of Cpn60.1 in *E. coli* including usage of different promoters, start codons, growth temperatures and strains which were not successful in achieving the required levels (Fan et al., 2012).

It was discussed in the introduction chapter (section 1.4.3) that tetradecameric structure and ATPase activity is necessary for a chaperonin to perform its chaperone activity (Bartolucci et al., 2005; Corrales & Fersht, 1996; Horwich & Fenton, 2009). However, the studies have showed that the *M. tuberculosis* chaperonins unusually exist as lower order oligomers and *M. tuberculosis* Cpn60.2 was demonstrated to crystalize

as a dimer (Goyal et al., 2006; Qamra et al., 2004; Qamra & Mande, 2004). The ability of the mycobacterial chaperonin Cpn60.2 to complement the loss of GroEL in *E. coli* strongly suggests that Cpn60.2 forms higher order oligomers *in vivo*. The study of Fan et al., 2012 showed that Cpn60.2 could form higher order oligomers under appropriate conditions though they are less stable than GroEL.

In another study, two Mtb*cpn60.2* plasmid constructs were used to compare the levels of expression required to complement the loss of GroEL in *E. coli*. One plasmid containing longer spacer region between Mtb*cpn10* and Mtb*cpn60.2* expressed the Mtb*cpn60.2* at lower levels compared to the expression levels of Mtb*cpn60.2* from a second plasmid containing a smaller spacer region. It was observed that only the Mtb*cpn60.2* expressed at higher levels could functionally complement the *E. coli* GroEL (Fan et al., 2012; Kumar et al., 2009a). This is consistent with (though it does not prove) the hypothesis that *in vivo* oligomerisation is needed for complementation. Oligomerisation is more likely to occur when the protein concentration is higher.

Another possible reason for the low expression levels of mycobacterial chaperonins in *E. coli* could be the codon bias, which is a universal phenomenon. In different organisms, because mostly of variation in GC content, the preferred codons vary. Due to high degree of redundancy in the genetic code, 64 genetic codons code for only 20 amino acids i.e., more than one codon encodes for the same amino acid. *E. coli* was the first unicellular organism in which the codon usage was studied and found to be strongly codon biased. Later, a positive correspondence was also observed between the frequency of the codons and the complementary tRNAs in the cells. When a heterologous protein is expressed in *E. coli*, ribosome could encounter a rare codon, pause, and eventually detach from the mRNA leading to reduced yield (Gutman & Hatfield, 1989; Quax et al., 2015; Rosano & Ceccarelli, 2009; Sharp, 2005). Researchers proposed the codon adaptation index (CAI) to measure the extent of the codon usage bias. CAI can be measured for a specific organism based on the measurement of the frequency of codon usage of a reference set of highly expressed genes in that organism. The CAI for a specific gene is measured by comparing the codon usage frequency of that gene with the reference set of highly expressed genes of that organism (Puigbò et al., 2008).

The complementation and expression analysis of *M. marinum* chaperonins has not yet been studied in detail. Therefore, the work described in this chapter and the next chapter (chapter 7) used different approaches to analyse the complementation and expression of chaperonins of *M. marinum* in *E. coli*. Different methods were also used to optimise the expression of *M. marinum* chaperonin genes in *E. coli*. It was hypothesized that *M. marinum cpn60.2* (*Mmcpn60.2*) would functionally replace GroEL in *E. coli* in combination with any of the co-chaperonins (*groES* and *Mmcpn10*) whereas, *M. marinum cpn60.1* (*Mmcpn60.1*) would not be able to replace the endogenous GroEL unless expressed at reasonably high levels in *E. coli*.

6.2 Experimental plan

For complementation and expression analysis of *M. marinum* chaperonins in *E. coli*, a high copy number expression plasmid pTrc99A (Figure 6.1) was used to construct the plasmid derivatives containing *M. marinum cpn60.1* (Mm*cpn60.1*) and *cpn60.2* (Mm*cpn60.2*) in combination with *groES* or *M. marinum cpn10* (Mm*cpn10*) under control of a tightly regulated *trc* promoter induced by IPTG or lactose.

Plasmid pTrc99A was chosen for complementation and expression analysis because previous complementation studies done in our lab had made extensive use of this plasmid. As this has a strong, controllable promoter, and to provide consistency with previous work, complementation plasmids constructed for the present study used the same backbone. Plasmids previously constructed in the lab were checked by sequencing and found to be derivatives of pBR322 containing either *groES*-Mm*cn60.1* or *groES*-Mm*cpn60.2*. These were therefore used as sources of fragments for construction of the relevant expression plasmids, as described in the section 6.3 and shown in Figures 6.4 and 6.5.

The complementation analysis was performed in two strains of *E. coli*: MGM100 (Figure 6.2) and AI90 (Figure 6.3). MGM100 is a derivative strain of *E. coli* K-12 MG1655 in which the *groE* promoter (regulating the *groESL* operon) has been replaced by *araBAD* promoter (p*BAD*) and *araC* regulatory gene. A kanamycin resistance cassette has also been inserted downstream of *groEL* (Figure 6.2) (McLennan & Masters, 1998; Nielsen et al., 1999). *E. coli* MGM100 strain grows well in the presence of arabinose but cannot grow in the presence of glucose because of repression of P_{BAD} promoter resulting in the depletion of essential GroESL (Guzman et al., 1995). AI90 is an *E. coli* strain in which the chromosomal copy of *groEL* gene (Δ*groEL*) has been replaced by kanamycin resistance cassette. As GroEL is essential for the survival of the cell, growth is maintained by *trc* regulated *groESL* on an expression plasmid. As the expression of *sacB* gene in *E. coli* presents sucrose sensitivity, growth of AI90 on sucrose allows the selection of cells which have lost the *sacB* expressing plasmid. This is only achievable by displacement of pACYC-*trc-groESL-sacB* plasmid with another plasmid carrying a functional chaperonin and co-chaperonin gene (plasmid shuffling) (Sivinski et al., 2021).

All plasmids used in the analysis contained chaperonin genes under control of *trc* promoter which was constructed by joining the -35 region of *trp* promoter and -10 region of the *lac* promoter, the resulting *trc* promoter is more tightly regulated than either of the two promoters (*trp* or *lac*) and can be induced by supplementing IPTG (or lactose) in the liquid or solid media (Tegel et al., 2011). Although *trc* promoter provides better expression levels than *lac* promoter, a small drawback is its high leaky expression of basal levels (Terpe, 2006). The advantage of using IPTG over lactose is that IPTG is not metabolized by the cells, so that its concentration remains constant enabling the continuous over-expression of genes on the plasmid under the control of the *trc* promoter.

Expression analysis was also performed in Tab21 strain of *E. coli* using SDS-PAGE. Tab21 strain was derived from BL21 (DE3) by a PhD student in our lab (Ahmed, 2010). BL21 (DE3) is the most widely used expression strain suitable for overproduction of proteins because it has T7 RNA polymerase gene under control of *lac* promotor (lactose / IPTG inducible) (Daegelen et al., 2009). To make the strain suitable for complementation and expression analysis of chaperonin genes, p*groE-groESL* operon of BL21 was replaced with p*BAD-groESL-kan*^{R} operon used in MGM100.

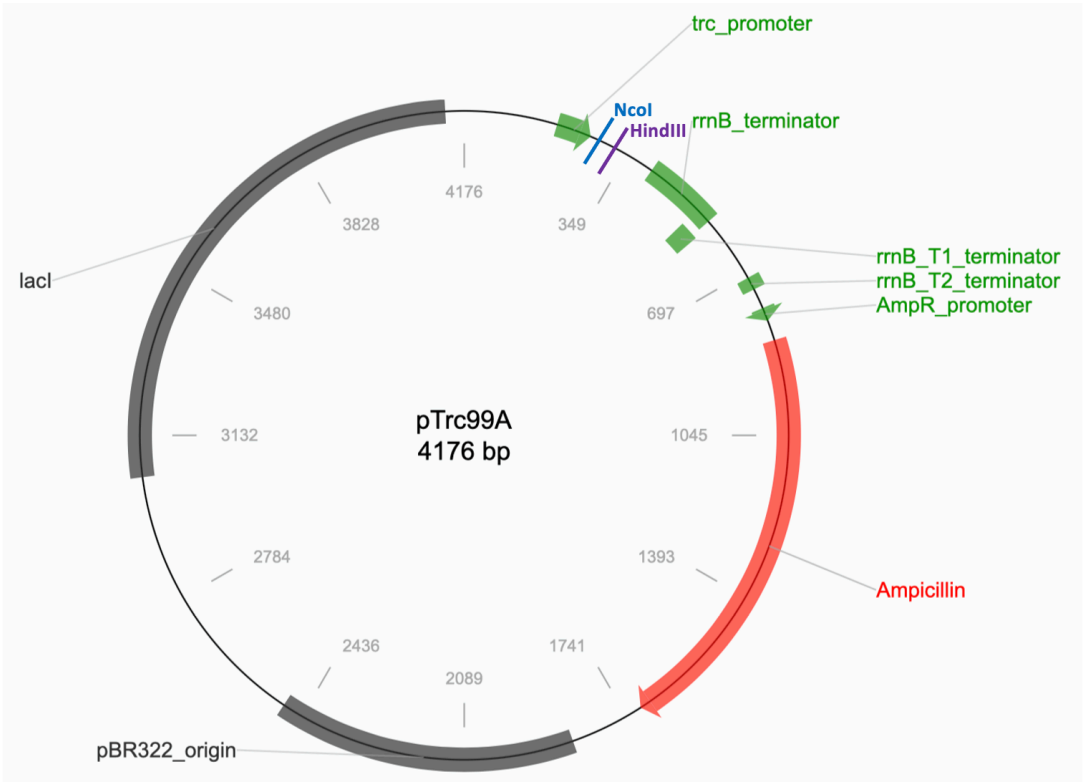


Figure 6.1 Map of cloning and expression vector pTrc99A.
Image adapted from https://www.addgene.org/vector-database/4402/. Restriction sites NcoI and HindIII are marked. Ampicillin resistance selection marker and *trc* promoter sites are also labelled.

.

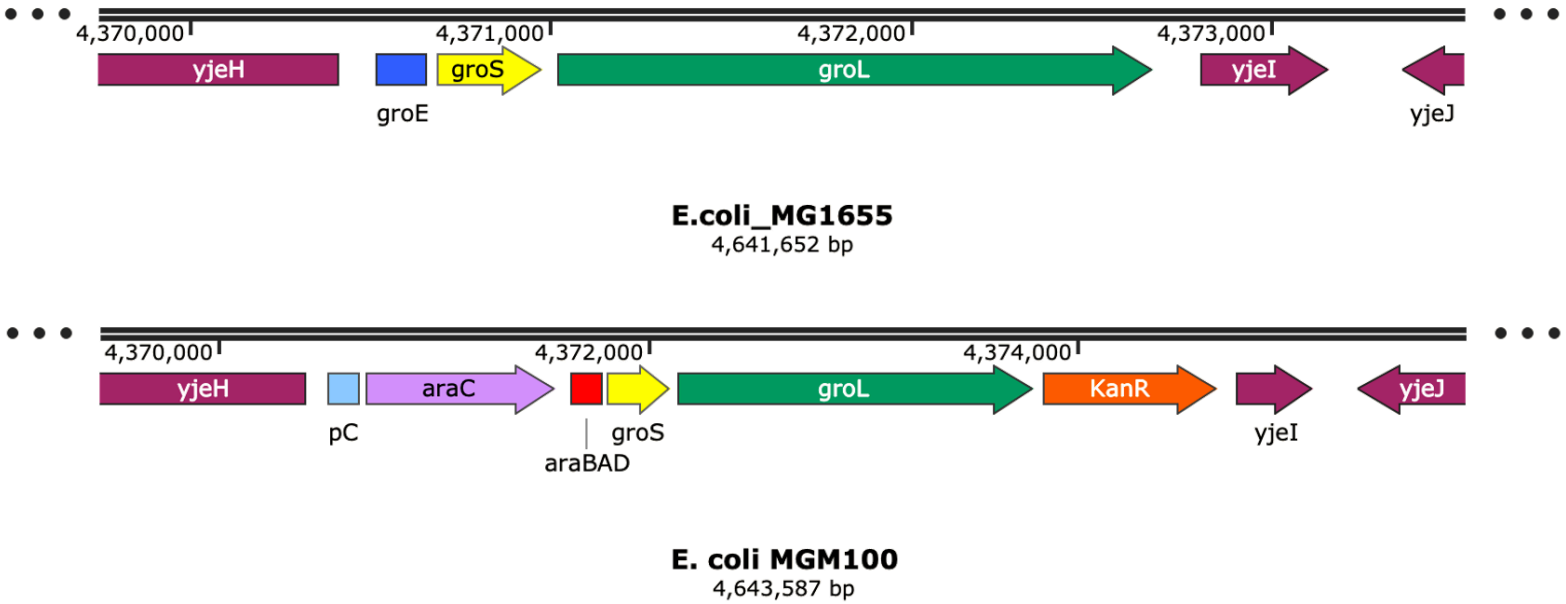


Figure 6.2 Genetic organization of *groESL* region and surrounding CDS of *E. coli* MG1655 and MGM100 strains.
In MGM100, the chromosomal copy of *groESL* can only be expressed in the presence of arabinose and in the absence of glucose. If there is glucose in the media, the chromosomal copy of *groESL* will be tightly repressed. In the presence of glucose and IPTG or in the absence of arabinose and presence of IPTG only the plasmid copy of the chaperonins will be expressed enabling the survival of the strains harbouring a plasmid construct with a functional combination of a chaperonin and a co-chaperonin. Therefore, MGM100 harbouring pTrc99A derivative plasmids carrying non-functional copy of a chaperonin and a co-chaperonin gene will not grow in the presence of glucose and IPTG.

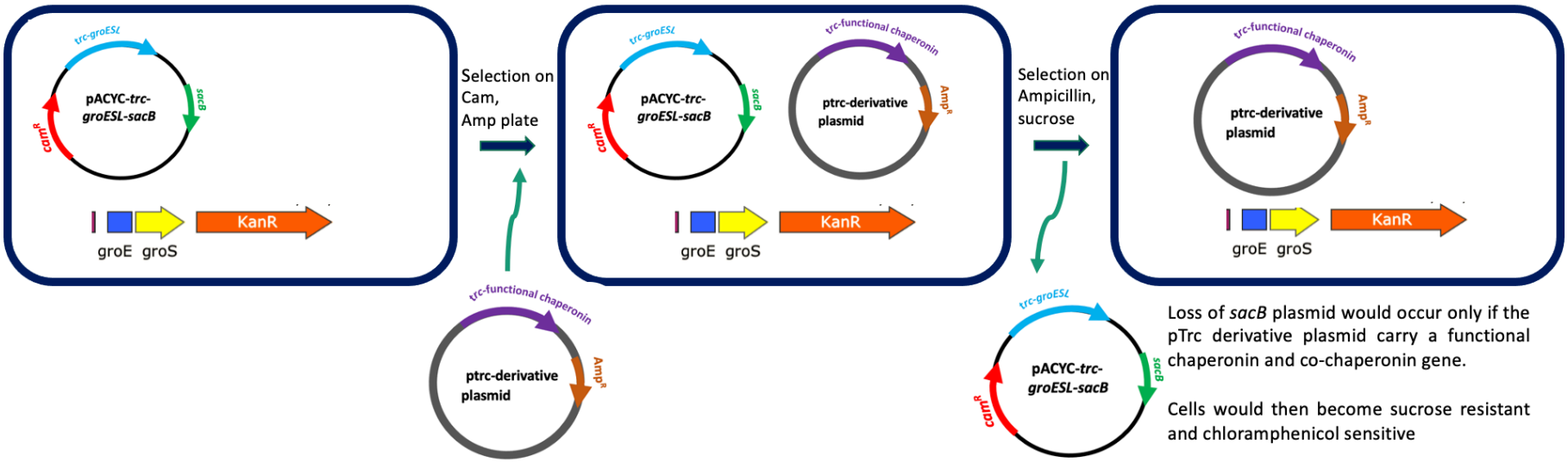


Figure 6.3 Schematic diagram showing the plasmid shuffling strategy in *E. coli* AI90 strain.

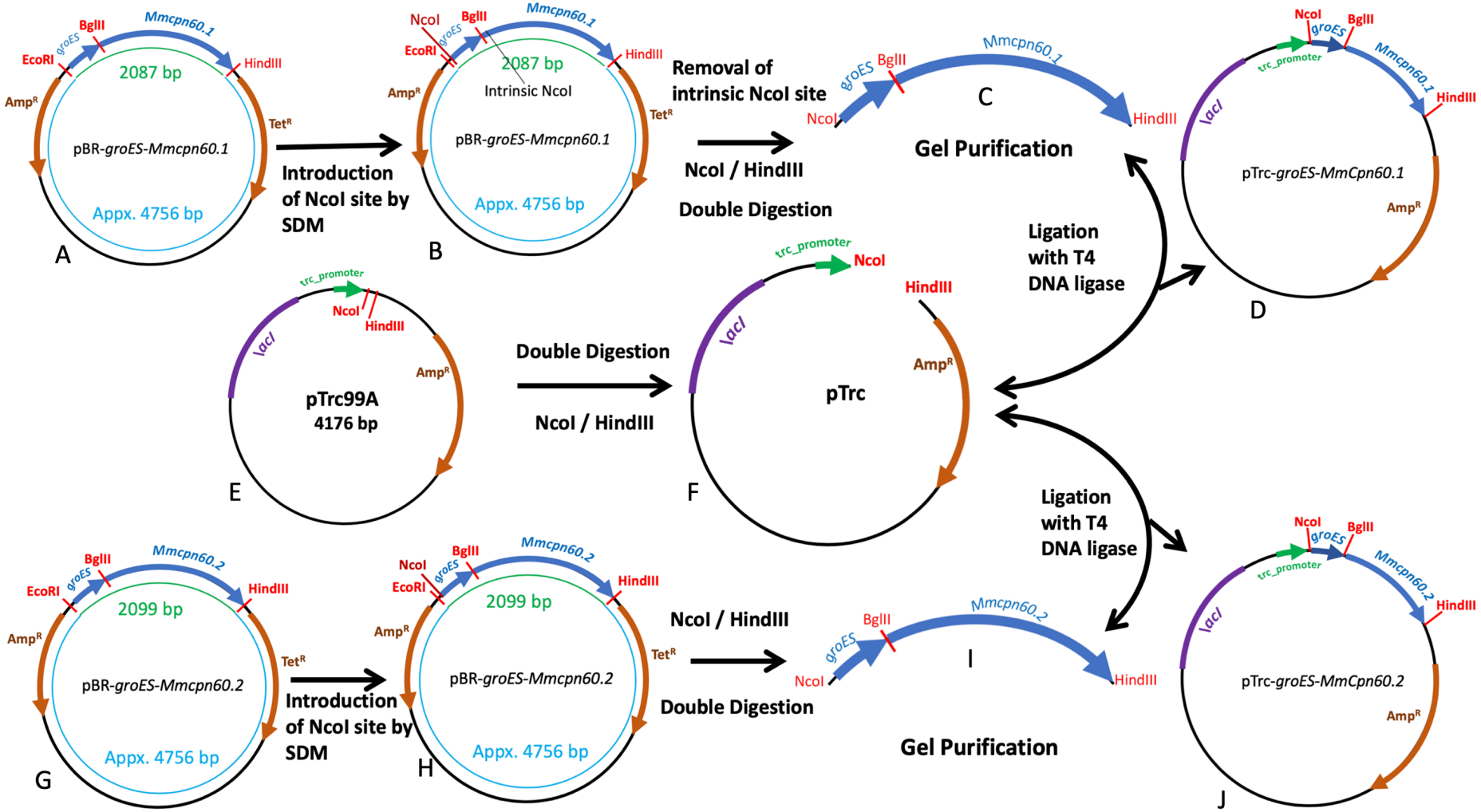


Figure 6.4 Schematic diagram of different steps of cloning of *groES*, Mm*cpn60.1*, and Mm*cpn60.2* into pTrc99A.

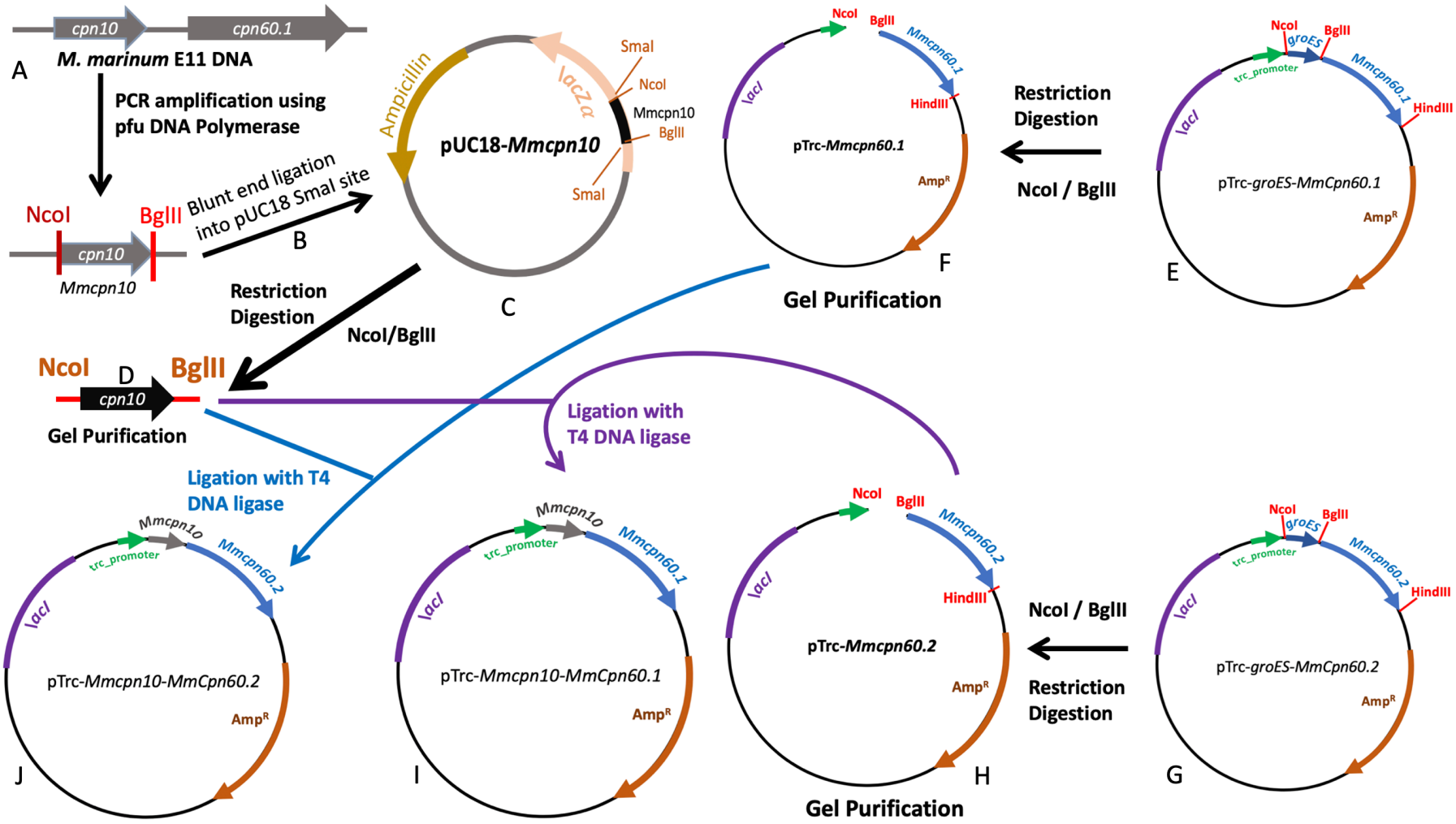


Figure 6.5 Schematic diagram of different steps of cloning of the Mm*cpn10*, Mm*cpn60.1* and Mm*cpn60.2* into pTrc99A.

6.3 Results

6.3.1 Construction of plasmids

For complementation and expression analysis of *M. marinum* chaperonins, pTrc99A derivative plasmids containing Mm*cpn60.1* and Mm*cpn60.2* in combination with *groES* and Mm*cpn10* under regulation of *trc* promoter were constructed as shown in the Figure 6.4 and Figure 6.5. For this purpose, two plasmid derivatives of pBR322, pBR-*groES*-Mm*cpn60.1* and pBR-*groES*-Mm*cpn60.2*, from the lab stock were used as a source of chaperonin genes fragments. pBR-*groES*-Mm*cpn60.1* contained *E. coli groES* and Mm*cpn60.1* cloned between EcoRI and HindIII restriction sites whereas, pBR-*groES*-Mm*cpn60.2* contained *E. coli groES* and Mm*cpn60.2* cloned between the same restriction sites. Both the plasmids had two resistance markers i.e., ampicillin and tetracycline (Figure 6.4 A & G).

6.3.1.1 Construction of pTrc-*groES*-Mm*cpn60.1* and pTrc-*groES*-Mm*cpn60.2*

In both the plasmids, NcoI restriction site was introduced just upstream to *groES* using site directed mutagenesis (SDM) following the protocol given in chapter 2 (section 2.6.3). It was later noticed that there was an intrinsic NcoI site in the gene sequence of *Mmcpn60.1* (40 bps downstream to the start codon). Therefore, the intrinsic NcoI site from the Mm*cpn60.1* gene in the plasmid pBR-*groES*-Mm*cpn60.1* was removed by a synonymous mutation converting GCC to GCA (for alanine) using site directed mutagenesis (Figure 6.4 B & H). Both the plasmids were then digested with NcoI and HindIII creating the fragments NcoI-*groES*-bglII-*Mmcpn60.2*-HinIII and NcoI-*groES*-bglII-*Mmcpn60.1*-HinIII which were gel purified and ligated into pTrc99A between NcoI and HindIII restriction sites (Figure 6.4). After ligation, the resulting plasmids were transformed into *E. coli* TOP10 for selection. The pTrc99A plasmid derivatives were first screened for the absence of tetracycline resistance (because pTrc99A carries only ampicillin resistant marker), then the clones carrying successfully inserted fragments were screened by restriction digestion and confirmed by sequencing. In this way, pTrc-*groES*-Mm*cpn60.1* and pTrc-*groES*-Mm*cpn60.2* were successfully

constructed. The protocols and primers used for SDM, restriction digestion, ligation and transformation are given in chapter 2.

6.3.1.2 Construction of pTrc-Mm*cpn10*-Mm*cpn60.1* and pTrc-Mm*cpn10*-Mm*cpn60.2*

The plasmids pTrc-*groES*-Mm*cpn60.1* and pTrc-*groES*-Mm*cpn60.2* were used to construct pTrc-Mm*cpn10*-Mm*cpn60.1* and pTrc-Mm*cpn10*-Mm*cpn60.2* by swapping *groES* with Mm*cpn10*. For this purpose, *M. marinum cpn10* gene was amplified from *M. marinum* DNA using polymerase chain reaction (PCR); pfu DNA polymerase enzyme was used in the PCR reaction (Figure 6.5-A). Following a PCR purification step, the blunt end fragment of Mm*cpn10* gene, having an NcoI and a BglII sites near the ends, was ligated into SmaI restriction site of pUC18 using the standard cloning and ligation conditions described in chapter 2, section 2.6.2 (Figure 6.5-B & C). The primers used for PCR amplification are given in chapter 2 (section 2.3) and the methods followed for PCR reaction set up and PCR purification are given in the section 2.6.2.2. Transformation of the ligation reaction was done in high efficiency *E. coli* TOP10 chemically competent cells. For blue/white selection, X-Gal, IPTG and ampicillin (selection marker for pUC18) were added in LB agar plates. Successfully cloned pUC18 derivatives containing Mm*cpn10* were screened by colony PCR of white colonies.

After successful cloning of Mm*cpn10* into pUC18, the plasmid was digested with NcoI and BglII restriction enzymes, the fragment NcoI-Mm*cpn10*-BglII was gel purified (Figure 6.5-D). The plasmids pTrc-*groES*-Mm*cpn60.1* and pTrc-*groES*-Mm*cpn60.2* were digested with same restriction enzymes to remove groES gene from both the plasmids. The plasmid backbone was gel purified after digestion (Figure 6.5-E, F, G & H). The fragment NcoI-Mm*cpn10*-BglII was ligated between NcoI and BglII sites in pTrc-Mm*cpn60.1* and pTrc-Mm*cpn60.2* backbone. All the plasmid constructs were verified by restriction digestion and sequencing to ensure the cloning of the required genes without any unwanted mutations.

6.3.2 Complementation assay in *E. coli* MGM100

To perform complementation assay in *E. coli* MGM100, expression plasmids constructed in sections 6.3.1.1 and 6.3.1.2 were transformed into *E. coli* MGM100 competent cells. MGM100 harbouring pTrc99A was used as a negative control whereas MGM100 strain carrying *pTrc-groESL* plasmid was used as a positive control. For comparison of the complementation analysis of *M. marinum* chaperonins with those of *M. smegmatis* and *M. tuberculosis*, pTrc99A derivative expression plasmids containing *M. smegmatis* chaperonin genes (Ms*cpn60.1* and Ms*cpn60.2*) and *M. tuberculosis* chaperonin genes (Mtb*cpn60.1* and Mtb*cpn60.2*) in combination with their respective co-chaperonins were also transformed into MGM100.

On 0.2% arabinose plate, all the strains showed optimal growth because arabinose induced the pBAD promoter and the chromosomal copy of *groESL* was expressed (Figure 6.6 A). On 0.2% glucose plate (causing tight repression of pBAD promoter), no strain was expected to grow because of the absence of IPTG. However, strains harbouring pTrc-*groESL* plasmid grew up to 10^{-6} dilution indicating the leaky expression of *trc* promoter and that the basal levels of GroES and GroEL expressed from the plasmid are enough to carry out the essential chaperonin activities (Figure 6.6 C). In addition, strain harbouring pTrc-Ms*cpn10-Mscpn60.2* also showed growth up to first three highest dilutions indicating that Cpn60.2 of *M. smegmatis* might be sufficient for complementation in smaller amounts compared to that of *M. tuberculosis* and *M. marinum* (Figure 6.6 C). Moreover, in the presence of 0.2% glucose along with 0.1mM IPTG, Cpn60.2 of *M. marinum* was able to complement the loss of GroEL in either combination with GroES or MmCpn10. However, the combination of MmCpn60.2 with GroES was found to complement better compared to MmCpn10 (Figure 6.6 D). Similarly, Ms*cpn60.2* and Mtb*cpn60.2* showed complementation in combination with Ms*cpn10* and Mtb*cpn10* respectively. On the other hand, *cpn60.1* of all the three species did not show complementation either with *groES* or *cpn10* of the respective species (Figure 6.6 D).

The concentration of 0.1mM IPTG was found to be the best working concentration for all strains because at higher concentrations cells harbouring pTrc-*groESL* or pTrc-Ms*cpn10-Mscpn60.2* plasmids showed restricted or no growth at all suggesting that the

higher levels of functional chaperonins might be lethal to the cells (Figure 6.7 A-D). Similar results were found on 0.2% arabinose and 0.1mM IPTG plate; both the chromosomal and plasmid copies of *groESL* were expressed resulting in the poor growth of cells (Figure 6.6 B, lane 2). A summary of the results is given in table 6.1 where +++ shows 60-100% growth while ++, + and – shows 20-50%, 1.0-10% and no growth respectively. Complementation assay was also performed in Tab21 at conditions similar to those used for MGM100. No difference was observed in the complementation results of any of the chaperonins.

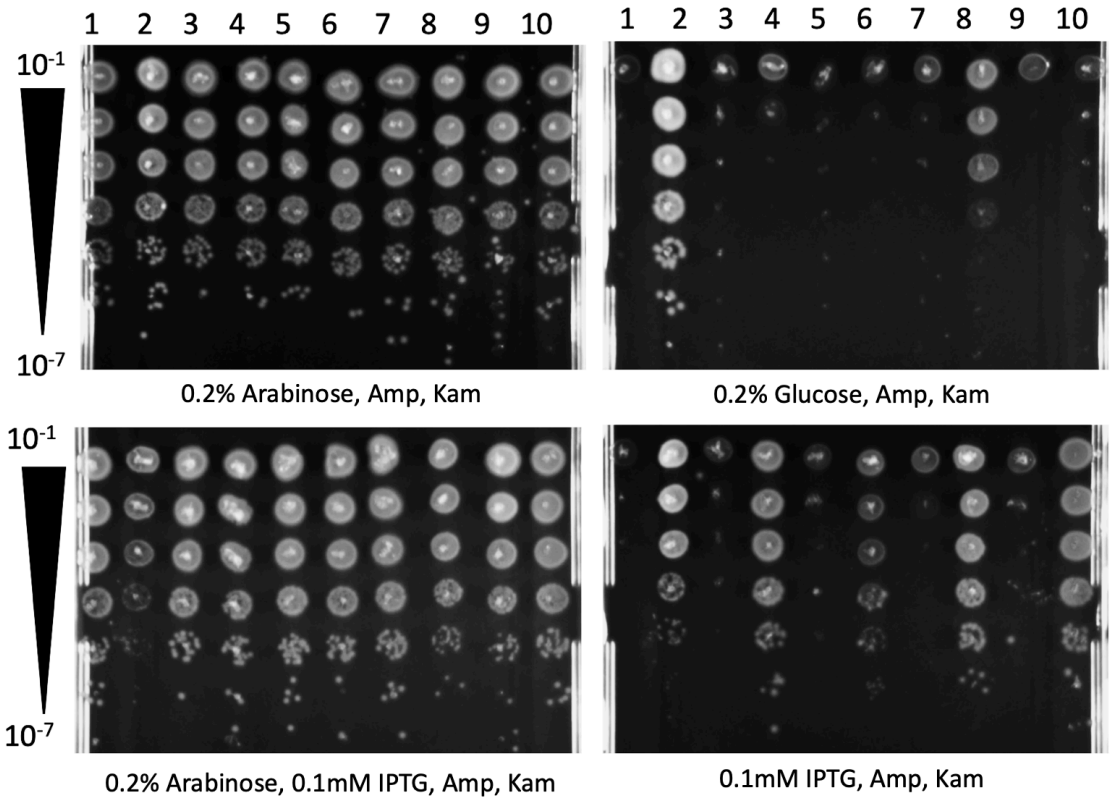


Figure 6.6 Complementation assay in MGM100 strain.
Complementation assay in MGM100 strain harbouring plasmids 1) pTrc99A (negative control), 2) pTrc-*groES-groEL* (Positive control), 3) pTrc-*groES*-Mm*cpn60.1*, 4) pTrc-*groES*-Mm*cpn60.2*, 5) pTrc-Mm*cpn10*-Mm*cpn60.1*, 6) pTrc-Mm*cpn10*-Mm*cpn60.2*, 7) pTrc-Ms*cpn10*-Ms*cpn60.*1, 8) pTrc-Ms*cpn10*-Ms*cpn60.2*, 9) pTrc-Mt*cpn10*-Mt*cpn60.1*, 10) pTrc-Mt*cpn10*-Mt*cpn60.2*. A single colony of each transformant was inoculated into 5 mL LB broth containing 0.2% arabinose, kanamycin (selection marker for MGM100) and ampicillin (selection marker for expression plasmid) and grown overnight in a shaking incubator at 30°C, 180rpm. The next morning each culture was pelleted down and washed twice with LBB followed by resuspension of the final pellet in 10 mL of LBB. OD_{600} of each culture was measured and normalized. In a 96-wells plate, ten-fold serial dilutions were made with the final dilution factor of 10^{-7}. 5 µl of each dilution was spotted on LB agar plates. The plates were incubated overnight at 30°C.

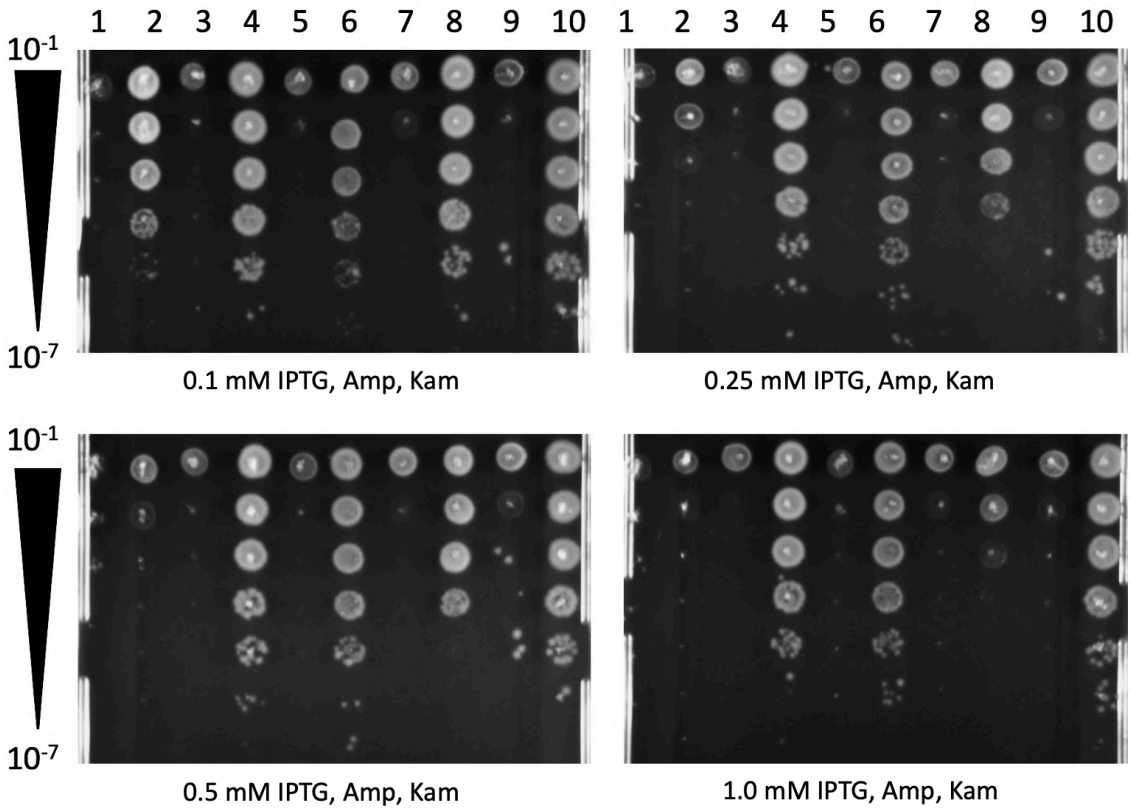


Figure 6.7 Complementation assay using MGM100 strain to find the best working concentration of IPTG.

Complementation assay using MGM100 strain harbouring plasmids to find out the optimum concentration of IPTG. 1) pTrc99A (negative control), 2) pTrc-*groES-groEL* (Positive control), 3) pTrc-*groES*-Mm*cpn60.1*, 4) pTrc-*groES*-Mm*cpn60.2*, 5) pTrc-Mm*cpn10*-Mm*cpn60.1*, 6) pTrc-Mm*cpn10*-Mm*cpn60.2*, 7) pTrc-Ms*cpn10*-Ms*cpn60.*1, 8) pTrc-Ms*cpn10*-Ms*cpn60.2*, 9) pTrc-Mt*cpn10*-Mt*cpn60.1*, 10) pTrc-Mt*cpn10*-Mt*cpn60.2*. Four plates with different concentrations of IPTG (0.1, 0.25, 0.5 and 1.0 mM) in the absence of arabinose (to shut down chromosomal copy) were also included in the analysis to confirm the optimum concentration required for induction. All plates contained kanamycin and ampicillin. The plates were incubated overnight at 30^{o}C.

Chaperonins	0.2% Arabinose	0.2% Arabinose + 0.1mM IPTG	0.2% Glucose	0.2% Glucose + 0.1mM IPTG	0.1mM IPTG	0.25mM IPTG	0.5mM IPTG	1.0mM IPTG
Negative control	+++	+++	-	-	-	-	-	-
GroES-GroEL	+++	++	+++	+++	+++	+	-	-
GroES-MmCpn60.1	+++	+++	-	-	-	-	-	-
GroES-MmCpn60.2	+++	+++	-	+++	+++	+++	+++	+++
MmCpn10-MmCpn60.1	+++	+++	-	-	-	-	-	-
MmCpn10-Mmcpn60.2	+++	+++	-	++	+++	+++	+++	+++
MsCpn10-MsCpn60.1	+++	+++	-	-	-	-	-	-
MsCpn10-MsCpn60.2	+++	+++	++	+++	+++	++	++	+
MtCpn10-MtCpn60.1	+++	+++	-	-	-	-	-	-
MtCpn10-MtCpn60.2	+++	+++	-	+++	+++	+++	+++	+++

Table 6.1 Summary of the complementation assay in *E. coli* MGM100.
Summary of the complementation assay in *E. coli* MGM100 harbouring expression plasmids carrying different combinations of chaperonins under control of *trc* promoter. Colour codes: light green (growth due to chromosomal GroESL), dark green (both the chromosomal and plasmid copies of the chaperonins expressed), red (leaky expression of *trc* promoter; basal levels of chaperonins are enough for survival), blue (restricted or no growth due to high levels of expression of the chaperonin, might be lethal to the cells), orange (chaperonin combination unable to complement the loss of GroEL), purple (chaperonin combination able to complement the loss of GroEL) and grey (low complementation).

6.3.3 Expression analysis using SDS-PAGE

The overall results of complementation assays with *E. coli* MGM100 demonstrated that Cpn60.2 of all the three mycobacterial species under study could complement the loss of native GroEL in either combination with GroES or Cpn10 of the respective species whereas Cpn60.1 failed to complement irrespective of the cochaperonin. The inability of Cpn60.1 to complement could either be due to the insufficient expression levels of Cpn60.1 or inability of Cpn60.1 to perform the essential chaperonin function which in turn could either be due to the inability of Cpn60.1 to recognize the GroEL substrate proteins or failure to fold them.

To estimate the expression levels of chaperonins, all the plasmids used in complementation assay 6.3.2 transformed in Tab21 were used. Cultures were grown and induced with IPTG. The proteins from the whole cell lysate were resolved using Sodium dodecyl-sulfate polyacrylamide gel electrophoresis according to the procedure given in chapter 2 section 2.7.1.

It was hypothesized that all the chaperonins expressed from the trc promoter would be visible on SDS-PAGE following induction with IPTG. However, the results of SDS-PAGE were unexpected; no distinguishable levels of any of the *M. marinum* chaperonin

(Mm*cpn60.2* & Mm*cpn60.1*) were seen. On the other hand, very clear bands of GroEL, MsCpn60.2 and MtbCpn60.2 were seen. No noticeable band of any of the mycobacterial Cpn60.1 weas observed on gels (Figure 6.8). It was examined in the complementation assay that *M. marinum* Cpn60.2 demonstrated the chaperonin activity in either combination with GroES or Cpn10 in the absence of a functional copy of GroESL (Figure 6.6 D & Figure 6.7). This raised the question, why MmCpn60.2 could not be over expressed to the detectable levels in Tab21. Moreover, we cannot rule out that Cpn60.1 of any of the three mycobacteria might not complement because of the lower expression levels.

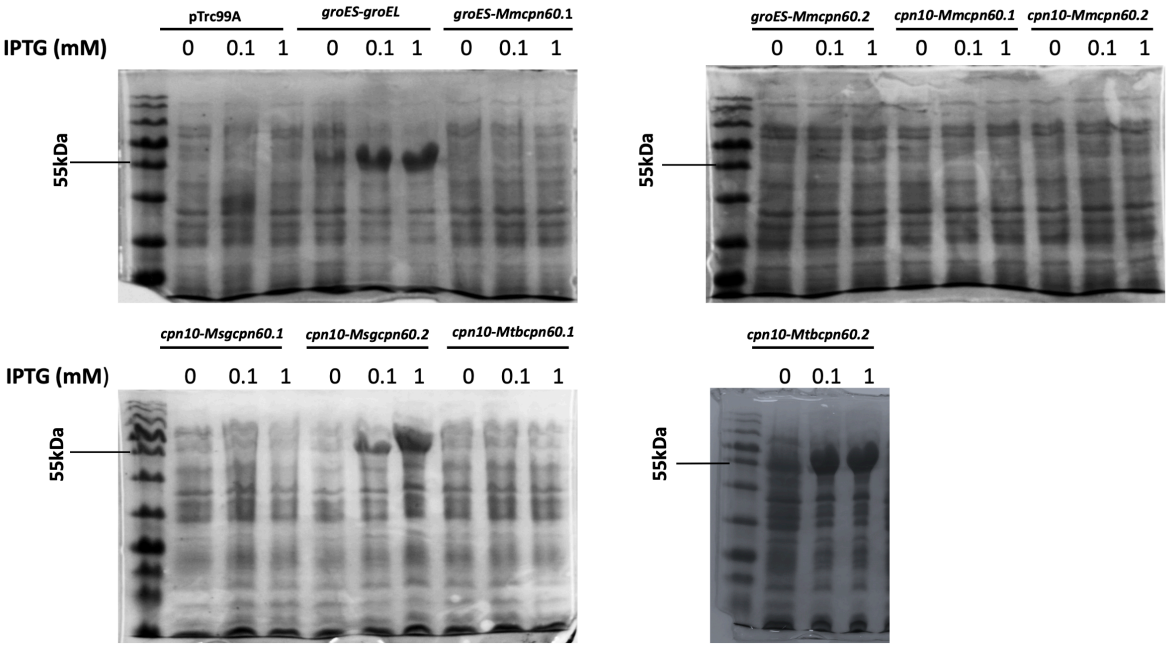


Figure 6.8 SDS-PAGE gel.

Images of the SDS-PAGE gels to resolve the proteins form the whole cell lysates of Tab21 strain harbouring pTrc derivative plasmids carrying different combinations of chaperonin genes under study. In the top left gel two over expression bands of GroEL (in the lanes 6 & 7) are clearly visible with 0.1 mM and 1mM IPTG, whereas in the bottom left gel two overexpression bands of *M. smegmatis* Cpn60.2 are visible (in the lanes 6 & 7) with the same concentrations of IPTG as above. Similarly, the bottom right gel is showing over expression bands of *M. tuberculosis* Cpn60.2. Unexpectedly, no band of any of the chaperonin of *M. marinum* was seen. One) pTrc99A (Negative control), two) pTrc-*groESL* (Positive control), three) pTrc-*groES*-Mm*cpn60.1*, four) pTrc-*groES*-Mm*cpn60.2*, five) pTrc-Mm*cpn10*-Mm*cpn60.1*, six) pTrc-Mm*cpn10*-Mm*cpn60.2*, seven) pTrc-Ms*cpn10*-Ms*cpn60.1*, eight) pTrc-Mm*cpn10*-Mm*cpn60.2*, nine) pTrc-Mt*cpn10*-Mt*cpn60.1*, ten) pTrc-Mt*cpn10*-Mt*cpn60.2*.

6.3.4 Complementation assay in *E. coli* AI90

Complementation assays were also performed in AI90 strain harbouring all the plasmids used for the complementation assay in MGM100. It was hypothesized that in AI90, pTrc plasmids carrying *cpn60.2* of all the three mycobacteria (i.e., *M. marinum*, *M. smegmatis* and *M. tuberculosis*) would be able to displace the plasmid pACYC-*trc-groESL-sacB* on 2% sucrose plate and would be able to rescue cell growth in the complete absence of GroEL.

In the presence of IPTG and absence of sucrose (plate A, Figure 6.9), all the cultures showed considerable growth either because of pACYC plasmid expressing the endogenous GroESL or the pTrc plasmid expressing complementing chaperonin and co-chaperonin. No strain showed growth on the plate containing 2% sucrose (plat B, Figure 6.9) except the cells harbouring pTrc-*groESL* plasmid. The reason for no growth was the loss of pACYC plasmid in the presence of sucrose and because of the absence of IPTG, pTrc plasmids could not express the chaperonin genes under control of *trc* promoter. The reason for the growth of the cells harbouring pTrc-*groESL* plasmid could be the leaky expression of trc promoter which again confirms that basal levels of endogenous GroES and GroEL are enough to sustain growth (confirming the results found in Figure 6.6, C).

Moreover, on the plate containing 2% sucrose along with IPTG only the cells have a functional copy of the chaperonins on the pTrc derivative plasmids were able to grow (plat D, Figure 6.9). Cells harbouring pTrc-Ms*cpn10*-Ms*cpn60.2* and pTrc-Mtb*cpn10*-Mtb*cpn60.2* grew well on 2% sucrose + IPTG plate demonstrating that *M. smegmatis* Cpn60.2 and *M. tuberculosis* Cpn60.2 could complement the complete loss of endogenous GroEL (plat D, Figure 6.9). Unexpectedly, *M. marinum* Cpn60.2 could not show complementation in AI90 which could be due to lower expression levels of MmCpn60.2 as seen on the SDS-PAGE gels (Figure 6.8). Cells harbouring *cpn60.1* (of all the three mycobacteria) carrying plasmids did not show any growth, indicated the inability of Cpn60.1 to complement the GroEL; which again could be due to insufficient expression levels.

We predict that a choramphenicol plus IPTG plate would allow growth of the Ms and Mtb Cpn60.2 clones. Chloramphenicol only plate confirmed the plasmid shuffling of

pACYC plasmid with pTrc-Ms*cpn10*-Ms*cpn60.2* and pTrc-Mtb*cpn10*-Mtb*cpn60.2* (lane 8 and 10). The growth in lane 2 (cells containing pTrc-*groESL*) could again be due to the leaky expression from *trc* promoter of pTrc-*groESL* (plate C, Figure 6.9). This again confirms that the basal levels of GroEL but not the MsCpn60.2 and MtbCpn60.2 were sufficient to sustain growth.

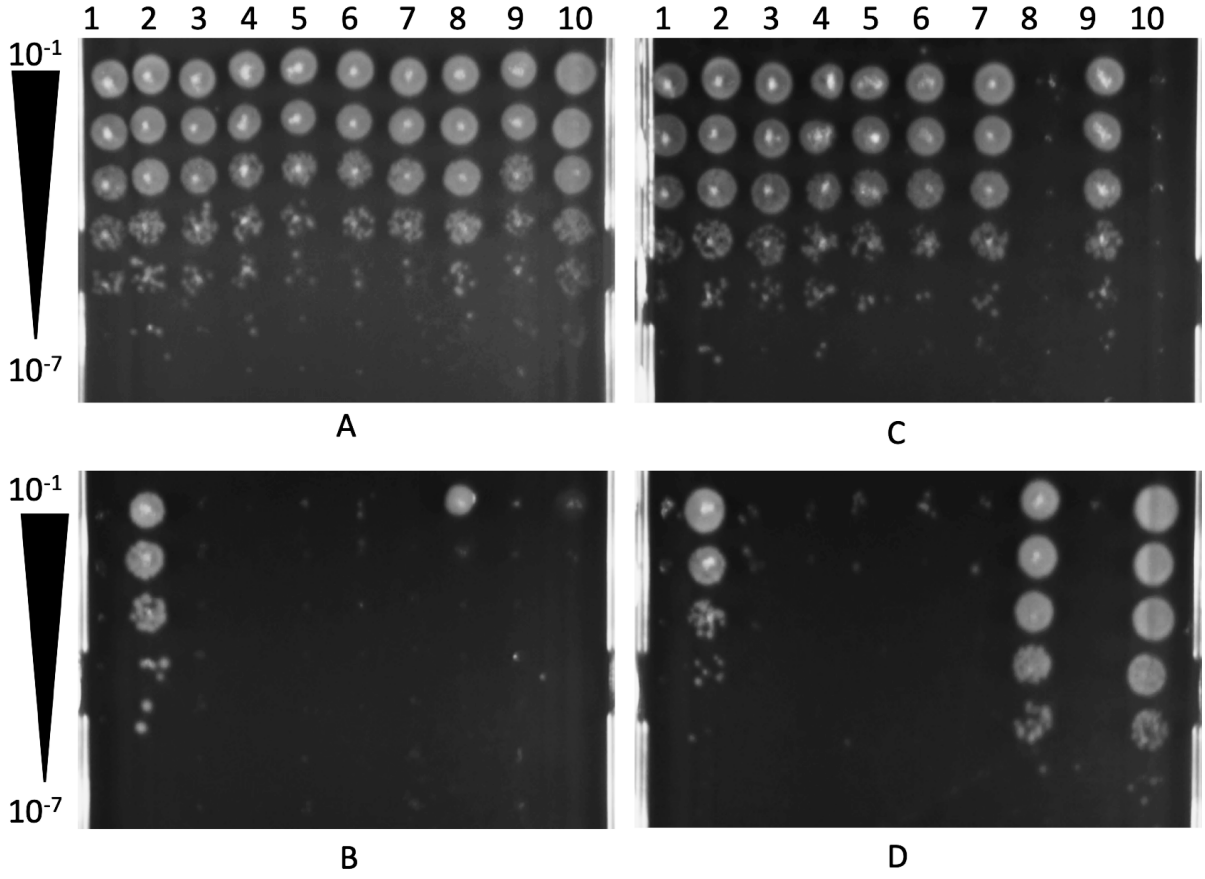


Figure 6.9 Complementation assay in *E. coli* AI90 strain.
(A) 0.5mM IPTG (B) 2% sucrose, (C) Chloramphenicol (D) 2% sucrose + 0.5mM IPTG. All plates contained ampicillin (selection marker for the pTrc plasmids) and kanamycin (selection marker for *E. coli* AI90 strain). Lanes 1) pTrc99A (negative control), 2) pTrc-groESL (Positive control), 3) pTrc-*groES*-Mm*cpn60.1*, 4) pTrc-*groES*-Mm*cpn60.2*, 5) pTrc-Mm*cpn10*-Mm*cpn60.1*, 6) pTrc-Mm*cpn10*-Mm*cpn60.2*, 7) pTrc-*Mscpn10-Mscpn60.1*, 8) pTrc-*Mscpn10-Mscpn60.2*, 9) pTrc-*Mtcpn10-Mtcpn60.1*, 10) pTrc-*Mtcpn10-Mtcpn60.2.* A single colony of each of the *E. coli* AI90 transformants was grown in 5 mL LB broth containing 1mM IPTG, kanamycin and ampicillin overnight at 30°C, 180rpm. The following day, 10 mL of the secondary growth cultures were set up from each primary culture and again grown overnight to give the cells enough time for plasmid shuffling. Next morning, OD_{600} of each culture was measured and normalized, 10-fold serial dilutions were made up to 10^{-7}, 5 µL of each dilution was spotted on selectable agar plates with or without sucrose.

The results of complementation assays (both in MGM100 and in AI90) and expression analysis indicated that *M. marinum* chaperonins behave differently in *E. coli* in comparison to the chaperonins of *M. smegmatis* and *M. tuberculosis*. Cpn60.1 did not show complementation phenotype which could be because of two possibilities. One is that Cpn60.1 does not function as a complete chaperonin under any circumstances. The other is that weak expression is the issue which might be due to the high G+C contents and codon usage bias or because of the formation of mRNA secondary structures, the *M. marinum* chaperonins were unable to be expressed at distinguishable levels in the *E. coli* cells. Secondly, the distance of the ribosome binding site (RBS) from the start codon of the *M. marinum* chaperonin genes was found to be strikingly shorter than that of *M. smegmatis* and *M. tuberculosis* chaperonin genes. And finally, the reason could be totally unexpected and unknown. To improve the expression of Mm*cpn60.1* in *E. coli*, codon optimized copy of the gene was synthesized from GenScript Biotech Corp (https://www.genscript.com) and used for complementation and expression analysis. RBS distance optimization was also performed as explained in the following sections.

6.3.5 Codon optimization of Mm*cpn60.1* using codon usage analysis

The expression levels of recombinantly expressed heterologous proteins are affected by codon usage bias. One possible reason for undetectable expression levels of *M. marinum* chaperonins in *E. col* could be the difference in codon usage of *M. marinum* compared to the expression host *E. coli.* Rare Codon Analysis tool by GenScript Biotech Corp (https://www.genscript.com/tools/rare-codon-analysis) was used to analyse the codon usage frequencies in terms of codon adaptation index (CAI) measurement for *M. marinum cpn60.1* and *cpn10* genes. The CAI values for Mm*cpn60.1* and Mm*cpn10* were found to be 0.60 and 0.65 respectively (Figure 6.10). A CAI value between 0.8 to 1.0 is considered to be good for expression of the desired heterologous protein in the host organism. For Mm*cpn60.1*, only 61% codons were found to have the usage frequency of 91-100 whereas 26% codons had less than 41% usage frequency out of which 5% had a usage frequency less than 11% (Figure 6.11-A).

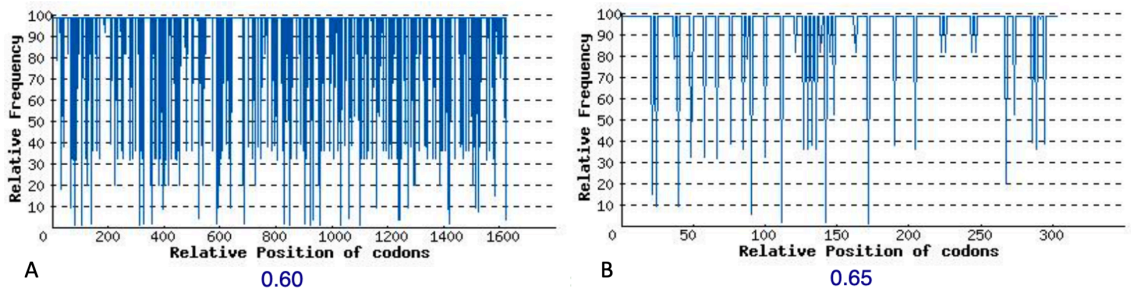


Figure 6.10 Graphs showing the distribution of codon usage frequency along the length of the gene sequences.
A) Mm*cpn60.1*, B) Mm*cpn10*.

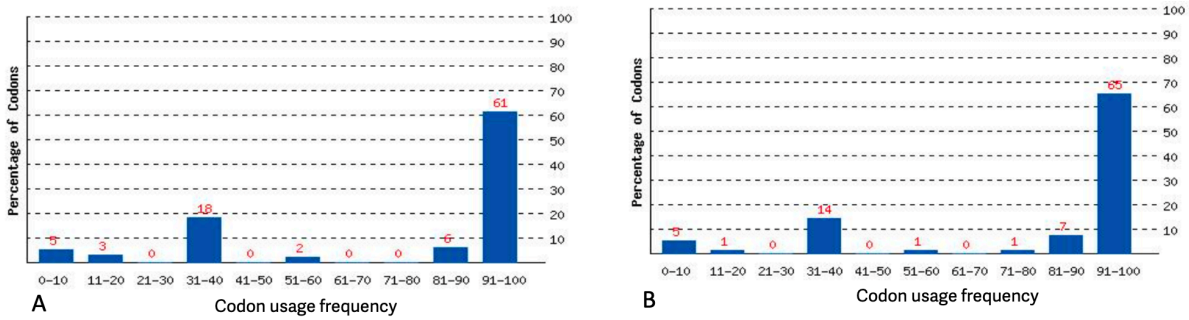


Figure 6.11 Percentage distribution of codons of Mm*cpn60.1* and Mm*cpn10*.
Graphs showing the percentage distribution of codons of A) Mm*cpn60.1* and B) Mm*cpn10* according to the codon usage frequency. A value of 100 represents the highest usage frequency for a given codon (encoding a given amino acid) in the host.

To optimize the CAI and the percentage of codons having highest usage frequency for the host organism (*E. coli*), Mm*cpn60.1* and Mm*cpn10* gene sequences were sent to GenScript Biotech Corp for codon optimization. The synthetic genes synthesized by GenScript Biotech had CAI upgraded from 0.60 to 0.97 for Mm*cpn60.1* and from 0.65 to 0.96 for Mm*cpn10* (Figure 6.12). Besides, tandem rare codons were removed and the percentage of codons having highest codon usage was improved from 61% to 87% for Mm*cpn60.1* and from 65% to 83% for Mm*cpn10*. All the codons with less than 70% usage frequency were replaced with the higher frequency codons (Figure 6.13). Moreover, GC content were also lowered from 67.72% to 61.20% for Mm*cpn60.1* and from 66.39% to 56.72% for Mm*cpn10*. Possible stem loop structures were also broken down to increase the stability and prolong the half-life of mRNAs.

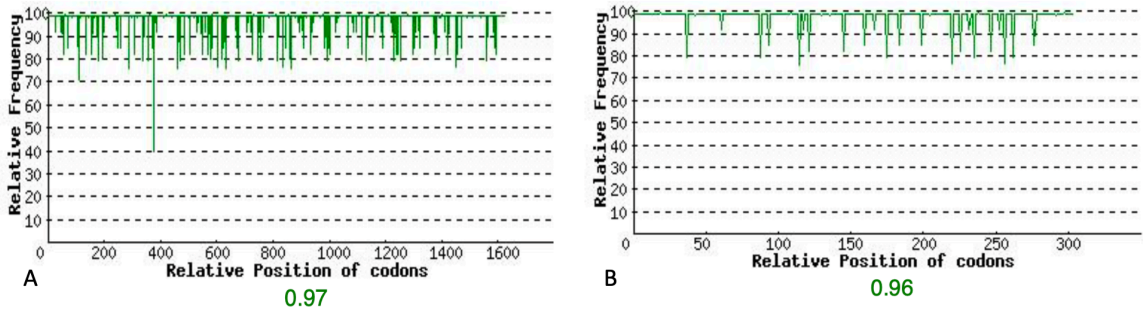


Figure 6.12 Percentage distribution of codons of codon optimized Mmcpn60.1 and Mmcpn10.
Graphs showing the distribution of codon usage frequency along the length of the gene sequences for the optimized sequences of A) Mm*cpn60.1*, B) Mm*cpn10*.

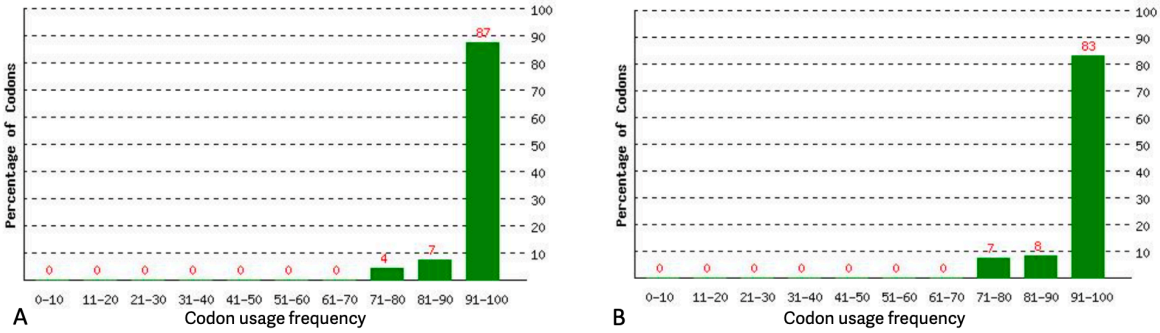


Figure 6.13 percentage distribution of codons of optimized gene sequences.
Graphs showing the percentage distribution of codons of optimized gene sequences of A) Mm*cpn60.1* and B) Mm*cpn10* according to the codon usage frequency. A value of 100 represents the highest usage frequency for a given codon (encoding a given amino acid) in the host.

6.3.5.1 Expression analysis of optimized Mm*cpn60.1* and Mm*cpn10*

Both the optimized genes of *M. marinum cpn60.1* and *cpn10* were received from the GenScript Biotech Corp harboured in the pTrc plasmid derivative under regulation of *trc* promoter (pTrc-Mm*cpn10*-Mm*cpn60.1*-CO). The expression plasmid was transformed into Tab21, and expression was analysed using SDS-PAGE. No detectable levels were observed on the SDS-PAGE gel.

6.3.6 Optimization of RBS distance from the start codon of Mm*cpn60.1*

In bacteria, ribosome binding site (RBS) is generally present 8 base pairs upstream to the start codon of a gene (Malys, 2012). In *M. marinum* the RBS distance from the start codon of Mm*cpn60.1* gene was found to be only 4 bps (Figure 6.14 A). The deviation from the optimum distance could potentially alter the strength of the RBS interaction with the ribosomes affecting the efficiency of translation leading to lower yields. The 4-bps spacer region was converted to 6, 8 and 10 bps using site directed mutagenesis (SDM). The primers and procedure followed for creating SDM are given in chapter 2, Table 2.7). The resulting plasmid constructs pTrc-Mm*cpn10*-Mm*cpn60.1* CO-6RBS, pTrc-Mm*cpn10*-Mm*cpn60.1* CO-8RBS and pTrc-Mm*cpn10*-Mm*cpn60.1* CO-10RBS were transformed into Tab21 strain and used for SDS-PAGE analysis. It was hypothesized that the expression of codon optimised Mm*cpn60.1* gene being expressed from the pTrc plasmids would increase to a distinguishable level upon induction with IPTG in any of the RBS distance optimized plasmids.

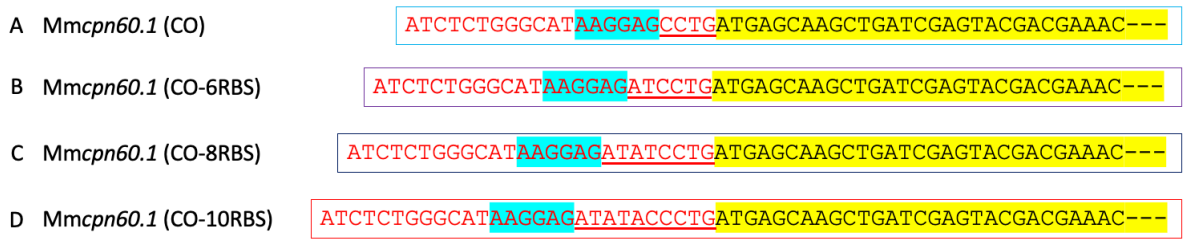


Figure 6.14 RBS distance optimization.
Mm*cpn60.1* gene sequence (yellow highlighted) and upstream region (red) showing a spacer region of only 4 bps between the start codon and RBS site in the W.T sequence (A). The 4-bps spacer region was changed to 6, 8 and 10 bps using SDM. Predicted RBS site is highlighted in turquoise.

6.3.6.1 Complementation and expression assay using sequence optimised and RBS optimised constructs

Complementation and expression assay was performed in Tab21 strain harbouring pTrc derivatives containing codon optimized and RBS distance optimized Mm*cpn10* and Mm*cpn60.1.* No complementation observed at 30°C (Figure 6.15) On SDS-PAGE gel no detectable levels of COMmCpn60.1 or CO plus RBS optimized were observed (Figure 6.16).

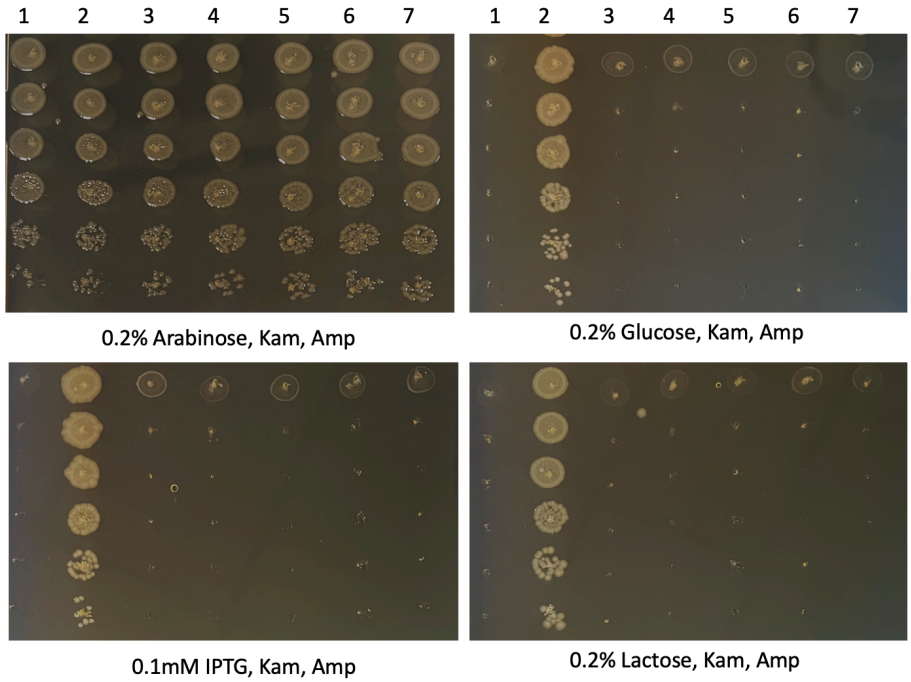


Figure 6.15 Complementation assay in Tab21 at 30°C.
1) pTrc99A (vector only control), 2) pTrc-gorES-groEL (positive control), 3) pTrc-Mm*cpn10*-Mm*cpn60.1* (original sequence), 4) pTrc-Mm*cpn10*-Mm*cpn60.1* CO (codon optimized) 5) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-6RBS 6) pTrc- Mm*cpn10*-Mm*cpn60.1*-CO-8RBS 7) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-10RBS. No complementation was observed for the first 24 hours of incubation at 30°C. The plates were left in the incubater for another night. Neither the CO nor the CO + RBS distance optimized MmCpn60.1 could show complementation for the loss of GroEL.

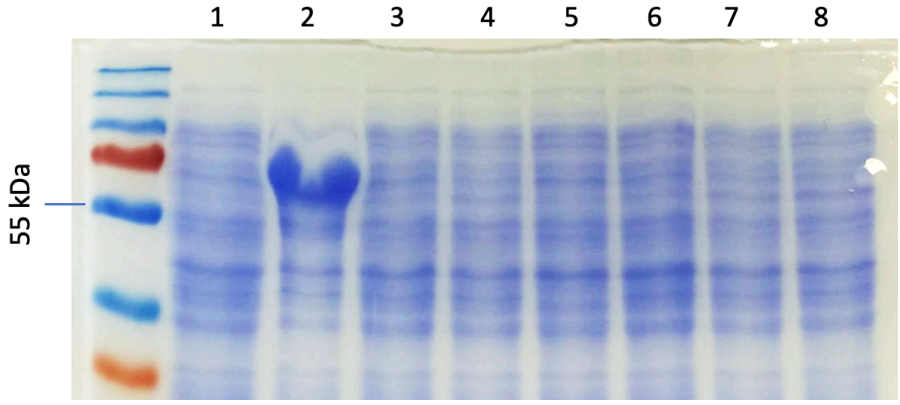


Figure 6.16 SDS-PAGE gel.
Expression analysis was performed in Tab21 harbouring plasmids induced with 1mM IPTG for 6 hours. 1) pTrc99A (vector only control), 2) pTrc-*gorES-groEL* (positive control), 3) pTrc-Mm*cpn10*-Mm*cpn60.1* (original sequence), 4) pTrc-Mm*cpn10*-Mm*cpn60.1* CO (codon optimized) 5) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-6RBS 6) pTrc- Mm*cpn10*-Mm*cpn60.1*-CO-8RBS 7) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-10RBS 8) Same as 7.

6.4 *E. coli* GroEL works as a functional chaperonin in combination with MmCpn10

Strong evidence is present for the role of chaperonins (both the Cpn60.1 and Cpn60.2) of *M. tuberculosis* in pathology of tuberculosis and granuloma formation. However, the immunological and cell-signalling properties of mycobacterial chaperonins are not fully understood (FRIEDLAND et al., 2008). Although Cpn60.1 is found to be a more potent cytokine stimulator than Cpn60.2, the immunogenic function of Cpn60.2 and its stimulating role in release of proinflammatory cytokines, such as TNF-α, from human monocytes in CD14 independent manner cannot be neglected (Lewthwaite et al., 2001). Joseph *et al*., demonstrated the anti-apoptotic role of Cpn60.2 showing that Cpn60.2 is exported out of the phagosomal membrane in *M. tuberculosis* infected macrophages and inhibits mitochondrial-dependent macrophage apoptosis (Joseph et al., 2017).

Experiments in our lab have validated the chaperonin function of Cpn60.2 by demonstrating that Cpn60.2 of *M. tuberculosis*, *M. smegmatis* and *M. marinum* can replace the loss of GroEL in *E. coli* (Ahmed, 2010; Rao, 2010, This study). Being essential for cell survival, *cpn60.2* deletion mutant cannot be constructed without expression of another gene that can play the essential chaperonin role; could it be *E. coli* GroEL? Can *E. coli* GroEL replace Cpn60.2 of *M. marinum*? We planned to construct a *M. marinum* Δ*cpn60.2* mutant by insertion of hygromycin resistant gene along with *E. coli groEL* in place of Mm*cpn60.2*. This mutant would only be possible to construct if GroEL can perform chaperonin function in combination with MmCpn10. To test if *E. coli* GroEL can work with *M. marinum* Cpn10, an expression plasmid containing *M. marinum cpn10* and *E. coli groEL* under control of *trc* promoter was constructed (pTrc-Mm*cpn10-groEL*, Figure 6.18).

It was proposed that *E. coli* GroEL would be able to function with *M. marinum* cpn10. The complementation assay in MGM100 showed that *E. coli* GroEL can perform the chaperonin function in combination with MmCpn10 (Figure 6.17). For construction of Mm*cpn60.2* KO mutant, codon optimized *groEL* sequence (to make it G+C rich and capable of being expressed in *M. marinum*) was obtained from GenScript. The plan was to infect zebrafish embryos with *M. marinum* Δ*cpn60.2* mutant and investigate its

impact on granuloma formation and inflammation but because of COVID-19 disruption this experiment was not performed.

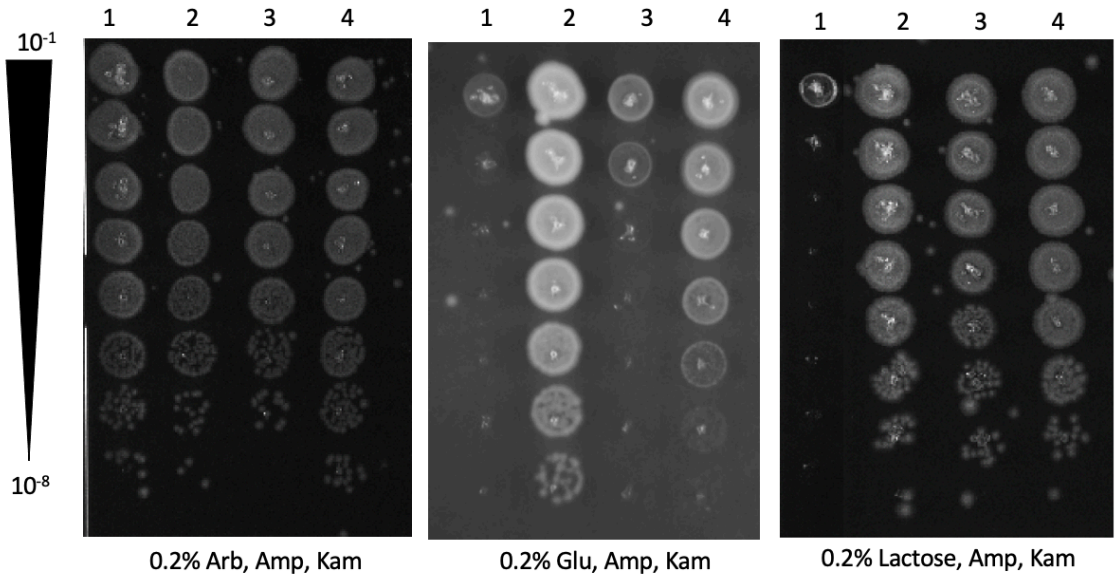


Figure 6.17 Complementation assay in MGM100 strain.
Complementation assay *in* MGM100 strain transformed with (1) pTrc99A (negative control vector), (2) pTrc-*groESL* (positive control), (3) pTrc-Mm*cpn10-groEL* (under question) and (4) pTrc-Ms*cpn10-groEL* (for comparison).
Chromosomal copy of *groESL* was suppressed by the absence of arabinose and presence of glucose in the agar plates while expression of the plasmid copy of chaperonins was induced by the addition of 0.1mM IPTG; therefore, only strains harbouring plasmids with the functioning chaperonins combination were able to grow on 0.2% glucose and 0.1% IPTG supplemented agar plate.

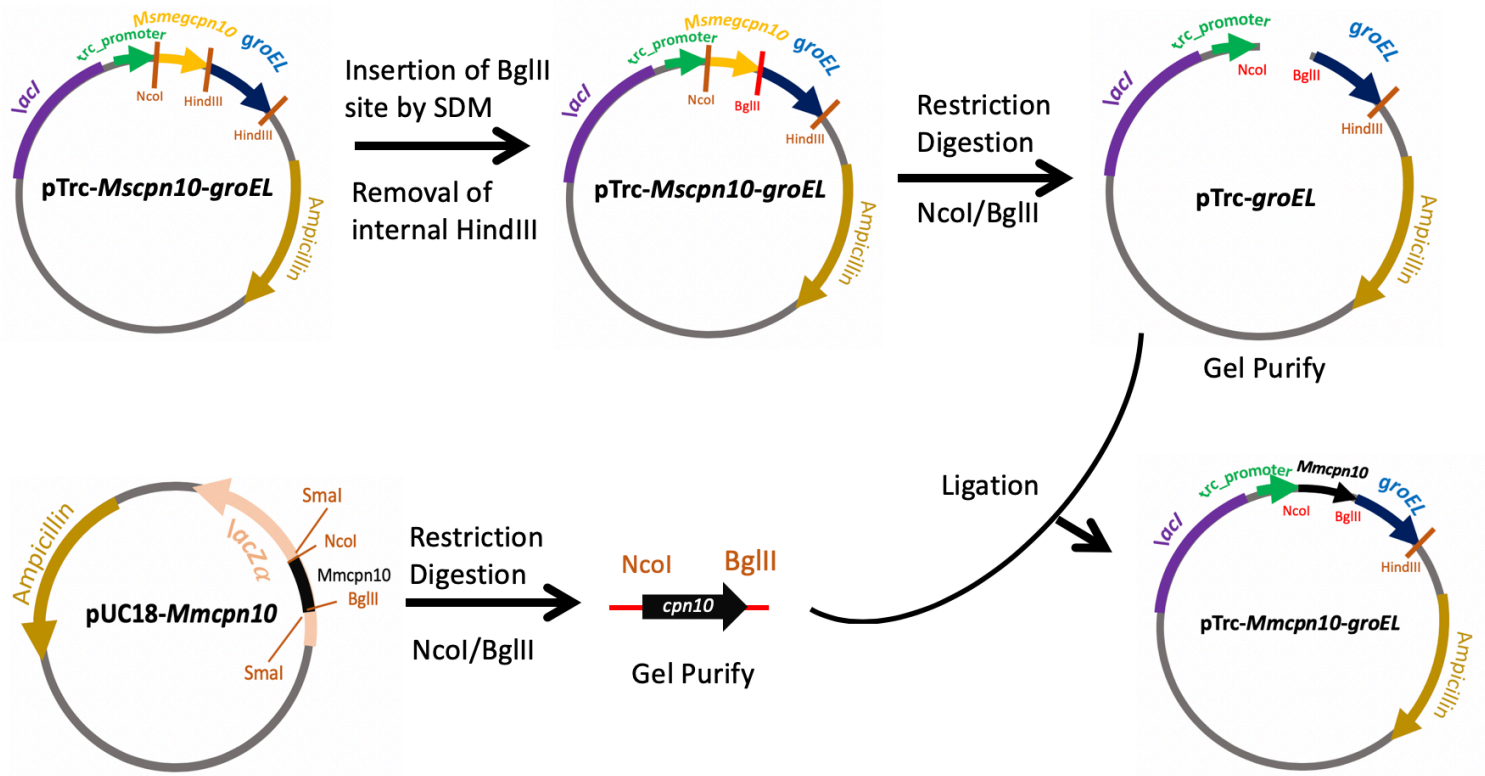


Figure 6.18 Schematic diagram of different steps of cloning for construction of pTrc-Mm*cpn10-groEL* plasmid.
The plasmid pTrc-*Mscpn10-GroEL* was obtained from the lab stock and sequenced using pTrc99A_F and pTrc99A_R prime. In this plasmid *Mscpn10-groEL* was cloned between NcoI and HindIII sites. In the intergenic region of *Mscpn10* and *groEL* was another HindIII site. The HindIII restriction site between *Mscpn10* and *groEL* was converted to BglII by site directed mutagenesis (Primers used for SDM are listed in Table 2.11). Introduction of BglII site was confirmed by double digestion and sequencing. The plasmid was then digested with NcoI and BglII. pTrc-*groEL* backbone purified by gel electrophoresis. *Mmcpn10* insertion fragment was gel purified after digestion of pUC18-*Mmcpn10* plasmid with the same restriction enzymes. The plasmid backbone pTrc-*groEL* and *Mmcpn10* insertion fragment were ligated overnight at 16°C.

6.5 Discussion

In this part of the study, complementation and expression analysis of *M. marinum* chaperonins was performed in *E. coli* using standard protocols. *M. tuberculosis* and *M. smegmatis* chaperonins were also included in the analysis for comparison. In *E. coli* MGM100 strain which has tightly repressed *groESL* operon, *M. marinum* Cpn60.2 was found to complement the loss of endogenous GroEL either in combination with GroES or Cpn10 of the respective species. Same results were found for the homologous Cpn60.2 from *M. tuberculosis* and *M. smegmatis*. However, the results were slightly different in AI90 strain in which MtbCpn60.2 and MsCpn60.2 were able to completely replace the function of the deleted chromosomal copy GroEL whereas, MmCpn60.2 was not found to be able to do so. In the complete absence of GroEL, strain expressing MmCpn60.2 was not able to sustain growth whether co-expressed with GroES or MmCpn10. One reason could be the lower expression levels of MmCpn60.2 (observed on the SDS-PAGE gel) compared to the MtbCpn60.2 and MsCpn60.2.

Despite several attempts, Cpn60.1 was not found to complement in any case neither in MGM100 nor in AI90. This again could be due to lower expression levels (observed on the SDS-PAGE gel) or a second reason could be the difference in the spectrum of the client proteins. RNA-seq analysis (chapter 5) suggested that Cpn60.1 might have some chaperonin activity in the cell because of which expression levels of Cpn60.2 were increased to many folds in the Δ*cpn60.1* strain to compensate the loss of chaperonin activity of Cpn60.1. If Cpn60.1 performs any chaperonin activity in *M. marinum* but does not show complementation in *E. coli*, this could be due to the difference in the spectrum of the client proteins of MmCpn60.1 and GroEL. Other reason could be oligomeric status of MmCpn60.1. The effect of oligomeric state on the chaperone activity is discussed in next chapter.

To improve the expression levels of MmCpn60.1 in *E. coli*, "sequence optimized" gene of Mm*cpn601* was used for expression and complementation analysis. Sequence optimization was done for codons usage, GC content and mRNA secondary structures. Ribosome binding site spacer region was also altered (optimized) in the optimised sequence plasmid to get the better expression. MmCpn60.1 optimised sequence with any of the optimised spacing of RBS from the start codon did not show better

expression (levels better than MmCpn60.2) and failed to complement in *E. coli* MGM100. Therefore, it was concluded that MmCpn60.1 was not able to fully function as a chaperonin when expressed in *E. coli*.

It was also observed that *E. coli* GroEL could perform the chaperone activity in combination with MmCpn10. It might be possible to create a *cpn60.2* KO mutant of *M. marinum* by replacing *cpn60.2* with codon optimized *groEL*. We planned to create this mutant for further testing the role of MmCpn60.2 in zebrafish infection. Due to COVID disruption this experiment was not performed. This could be an interesting idea for future researchers.

7 Further analysis of MmCpn60.1 function by complementation in *E. coli*

7.1 Introduction

It was demonstrated in chapter 6 that none of the Cpn60.1 proteins of the three mycobacterial species used in the complementation analysis (i.e., *M. tuberculosis, M. smegmatis* and *M. marinum*) showed functional chaperonin activity in *E. coli*. A few possible reasons for this were given in the discussion section of chapter 6. In this chapter, previous studies done to investigate the more possible reasons and further experiments done to test the possibilities will be discussed.

The structural and functional properties of mycobacterial chaperonins are not fully understood yet. Scientists use *E. coli* GroEL and GroES chaperonin and co-chaperonin as a prototypical example to explore and understand the mechanism and function of mycobacterial chaperonins. The detailed structure and protein folding mechanism of GroEL has been discussed in introduction chapter (section 1.3.7). *E. coli* chaperonin GroEL forms a tetradecamer of two heptameric rings and its cognate co-chaperonin GroES forms the lid for one of the two heptameric cavities (Hayer-Hartl et al., 1995; Xu et al., 1997b). The tetradecameric architecture and ATP hydrolysis is the fundamental requirement for a molecular chaperonin to perform its chaperone function of protein folding (Bukau & Horwich, 1998; Horwich et al., 2009; Horwich & Fenton, 2020; Rye et al., 1997; Xu & Sigler, 1998). However, repeated studies have reported that the mycobacterial chaperonins did not form the classical tetradecamer assembly of GroEL *in vitro* rather were crystalized as lower order oligomers and were found to exhibit weak ATPase activity (Qamra et al., 2004; Qamra & Mande, 2004). *M. tuberculosis* Cpn60.2 was crystalized as a dimer however was demonstrated to form higher order oligomers under unique environment conditions of salts and nucleotides (Fan et al., 2012; Qamra & Mande, 2004). The complementation ability of mycobacterial Cpn60.2 implies that it forms tetradecameric oligomer in vivo.

It was demonstrated in the previous chapter that the wild type Cpn60.1 of both the *M. tuberculosis* and *M. marinum* failed to show the chaperonin activity in *E. coli*. One possibility could be the lower order oligomerization state of mycobacterial Cpn60.1 in *E. coli* as hypothesized in the discussion section of chapter 6. It was also suggested that oligomerization of *M. tuberculosis* Cpn60.1 (MtbCpn60.1) in its indigenous cell environment is facilitated by post-translational modifications such as phosphorylation

which might require specific kinases; this could also be the reason of loss of its higher order oligomerization when heterogeneously expressed in *E. coli* (Kumar et al., 2009b). Qamra et al., 2004 demonstrated that many highly conserved residues in *E. coli* GroEL and Cpn60.2 of various mycobacterial species were found to be different in Cpn60.1 of *M. tuberculosis*. For instance, glutamic acid present at position 76 in *E. coli* GroEL and position 75 in MtbCpn60.2 was replaced by serine in MtbCpn60.1. Another example of such replacement is alanine at position 2 of *E. coli* GroEL and MtbCpn60.2 which was demonstrated to be crucial for the stabilization of tetradecameric structure of GroEL (Horovitz et al., 1993) was also found to be replaced by serine in MtbCpn60.1 (Qamra et al., 2004). Mutation of alanine (A) to serine (S) at position 2 (A2S) or glutamic acid (E) to serine (E76S) at position 76 in *E. coli* GroEL resulted in the destabilization of the tetradecamer structure (Qamra et al., 2004). Several other important residues in the equatorial domains were also found to be replaced by serine or threonine in MtbCpn60.1. This raised the possibility that some of them could be phosphorylated in vitro, and this might be required for full function as a chaperone (Kumar et al., 2009b).

Phosphorylation of proteins is one of the important mechanism an organism adopts in response to extracellular signals. The function of protein can be modified or become activated/inactivated because of alteration in the structural conformation caused by phosphorylation. Phosphorylation of proteins is a reversible post-translational modification process in which amino acids are phosphorylated by covalently bond phosphate groups catalysed by kinases (Cohen, 2002). In bacteria, one trigger for phosphorylation of proteins is as a response to the changes in environment. Bacteria use two component system consisting of histidine protein kinases and the associated regulators for phosphorylation of proteins (Stock et al., 1989b). Contrarily, in eukaryotes proteins are phosphorylated at serine/threonine or tyrosine amino acids by eukaryotic protein kinases. *M. tuberculosis* genome contains 11 genes encoding unique eukaryotic-like serine/threonine protein kinases (STPKs) (Av-Gay & Everett, 2000). Several studies have also shown the key role of these eukaryotic-like STPKs in various cellular process (such as stress response) in mycobacteria (Narayan et al., 2007; Wehenkel et al., 2008).

M. tuberculosis Cpn60.1 was found to exist in various lower order oligomeric forms such as monomeric, dimeric, heptameric and the assembly from heptameric to double

ring tetradecameric structure was found to be dependent upon phosphorylation on serine residues (Kumar et al., 2009b; Qamra et al., 2004). Immunoprecipitation followed by immunoblotting showed that the tetradecameric MtbCpn60.1 was phosphorylated whereas the heptameric form was not phosphorylated. Mass spectrometry analysis confirmed the phosphorylation at serine-393 position in the tetradecameric MtbCpn60.1 (Kumar et al., 2009b).

In further work conducted in our lab, Dr CM Santosh Kumar has constructed various single and double phosphomimetic mutants of MtbCpn60.1 in our lab by converting serine (S) at positions 75, 393, 416, 444 and 484 to aspartic acid (D) and investigated the chaperonin activity of the mutants by complementation analysis in *E. coli* MGM100. It was observed that no strain harbouring plasmid containing single or double mutant MtbCpn60.1 was able to complement the loss of GroEL (suppressed) in MGM100 except one double mutant S75D, S393D (which we refer to as MtbCpn60.1-DPM) which sustained cell growth in the absence of GroEL (Figure 7.1). However, MtbCpn60.1 DPM did not show chaperonin function in the *groEL* KO strain AI90 which suggests that either the protein is not fully functional, or it requires some residual GroEL to perform its full chaperonin activity.

As we know that ATPase activity is necessary for a chaperonin to perform its function of protein folding. It was hypothesized that the ATPase mutation would abolish the chaperone activity of MtbCpn60.1-DPM. To test this hypothesis, the double phosphomimetic mutant was further mutated at the ATP binding amino acids. For this purpose, three aspartic acid (D) residues at positions 51, 86 and 396 (corresponding to 52, 87 and 398 positions of *E. coli* GroEL respectively) were mutated to alanine to eliminate the ATPase activity of MtbCpn60.1 DPM. The complementation assay using ATPase mutants showed that all the three mutations abolished the complementation ability of MtbCpn60.1 DPM (Figure 7.2). These experiments suggest that the double phosphomimetic mutant of MtbCpn60.1 DPM might be able to form a tetradecamer structure which allowed the protein to replace the GroEL function.

S75 and S393 might not be the sole phosphorylation sites of MtbCpn60.1 as shown by (Canova et al., 2009). In vitro phosphorylation of MtbCpn60.1 using seven Ser/Thr protein kinases (STPKs) demonstrated that all the seven kinases were able to

phosphorylate but PknF produced the strongest signal and showed the phosphorylation of MtbCpn60.1 at two threonine positions i.e., T25 and T54 (Canova et al., 2009). The residues corresponding to T25 and T54 of MtbCpn60.1 are alanine (A) and serine (S) in GroEL and Cpn60.2 (Figure 7.3). Also, we don't know whether the phosphomimetic mutations at these sites would be able to make the protein fully functional as a chaperone. The same kinase enzyme PknF could not phosphorylate *M. smegmatis* Cpn60.1 (Canova et al., 2009) indicating that phosphorylation mechanism might be different in different mycobacteria.

I could not find information about the phosphorylation of MmCpn60.1 in the published literature. In chapter 6, it was shown that the expression levels of *M. marinum* Cpn60.1 in *E. coli* did not increase to detectable level on SDS-PAGE by sequence optimization (CO) followed by optimization of the RBS spacer region from the start codon. In this part of the study, various phosphomimetic mutants of MmCpn60.1-CO were constructed and used for expression and complementation analysis to test the hypothesis that phosphorylation at the same residues (as shown by Dr Santosh for MtbCpn60.1-DDPM) in MmCpn60.1 might cause the protein to oligomerize and perform the chaperonin function in *E. coli*.

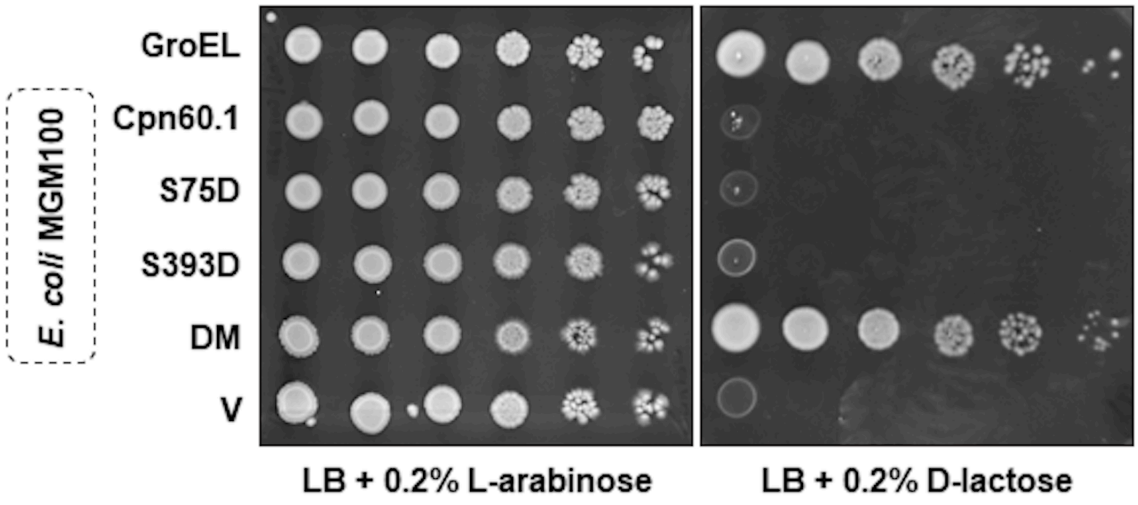


Figure 7.1 Complementation analysis of single and double phosphomimetic mutants of MtbCpn60.1.
Complementation analysis (at 30°C) in MGM100 harbouring plasmid containing IPTG or lactose inducible MtbCpn60.1 single and double phosphomimetic mutants i.e., S75D, D393D and DM (S75D+S393D). Image taken from Dr CM Santosh Kumar.

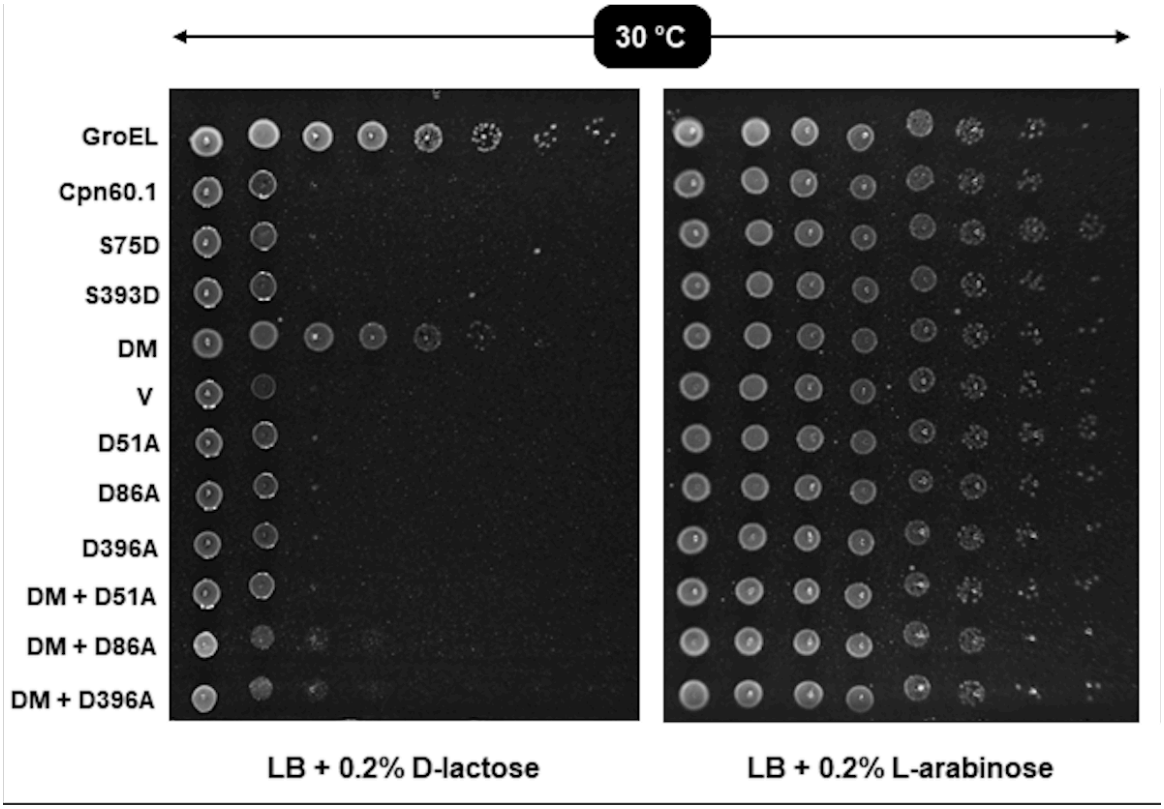


Figure 7.2 Complementation analysis of double phosphomimetic plus ATPase mutants of MtbCpn60.1.
Complementation assay in MGM100 harbouring plasmids containing IPTG or lactose inducible MtbCpn60.1 phosphomimetic mutants and ATPase mutants. Image taken from Dr CM Santosh Kumar.

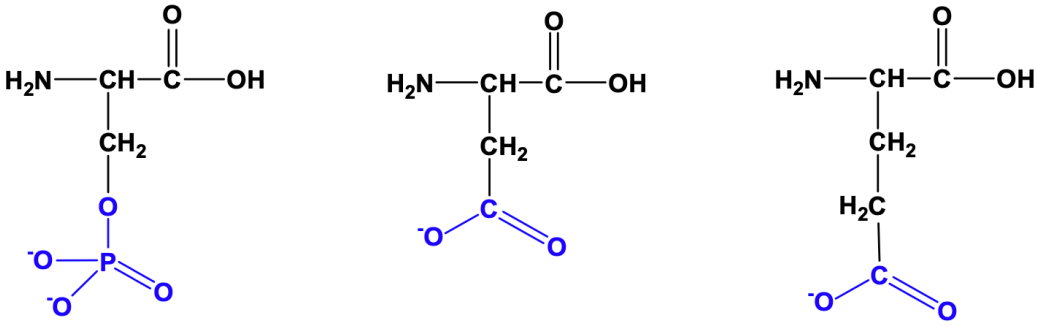
7.2 Experimental plan

The experiments of Dr CM Santosh Kumar in our lab with MtbCpn60.1 phosphomimetic mutants demonstrated that a double phosphomimetic mutant (S75D+S393DD) of MtbCpn60.1 was able to function as a chaperonin in MGM100 by enabling the cell growth in repressed GroEL environment at 30°C. To investigate whether the phosphomimetic mutations of *M. marinum* Cpn60.1 at the same positions would complement the loss of GroEL in *E. coli*, we chose serine and threonine residues of MmCpn60.1 at the positions corresponding to the positions tested by Dr Santosh in MtbCpn60.1. For this purpose, the protein sequences of Cpn60.1 and Cpn60.2 from ten mycobacterial species were aligned with GroEL to investigate the serine/threonine positions in Cpn60.1 proteins and to look for the conservation of those position among different mycobacterial species.

The multiple sequence alignment showed that at serine (S) residue was present at position 75 of Cpn60.1 of nine out of ten mycobacterial species included in the analysis; only *M. lepromatosis* Cpn60.1 have alanine (A) instead of serine at this position. *E. coli* GroEL and Cpn60.2 of all the ten mycobacteria had a highly conserved glutamic acid (E) at the corresponding position (Figure 7.3). At position 393 of Cpn60.1 serine was present in 8 out of 10 mycobacterial species whereas *E. coli* GroEL and Cpn60.2 of all the ten mycobacterial species had a highly conserved arginine (R) at the corresponding position (Figure 7.4).

M. tuberculosis genome encodes for 11 serine/threonine protein kinases (STPKs). In one study *in vitro* phosphorylation of MtbCpn60.1 with various STP kinases was performed. Different kinases were demonstrated to cause phosphorylation of MtbCpn60.1 but PknF produced the strongest signals for *in vitro* phosphorylation of MtbCpn60.1 (Av-Gay & Everett, 2000; Canova et al., 2009). All the STPKs of *M. marinum* are not studied yet therefore, the genome of 1218R was searched to look for the protein kinases homologous to the STPKs of *M. tuberculosis* and the results are presented in the next section 7.3.1. Multiple sequence alignment of the STPKs of both the species was performed to find the *M. marinum* STP kinases potentially homologous to the kinases of *M. tuberculosis*. It was hypothesized that if both the species have similar kinases, they might have similar mechanism and sites of phosphorylation of Cpn60.1.

For making phosphomimetic mutants, I chose two serine (S) residues at positions 75 and 393 of MmCpn60.1 and converted them to aspartic acid (D). The single and double phosphomimetic mutants were created and tested for complementation and expression in *E. coli* strains. The reason for converting serine to aspartic acid is that aspartic acid (like glutamic acid) can mimic a phosphorylated amino acid (such as phosphoserine or phosphothreonine) (structures below). If an enzyme becomes activated upon phosphorylation at serine/threonine or tyrosine residues, mutating them to aspartic acid or glutamic acid may cause the enzyme to remain in activated position permanently.



Structural comparison of aspartic acid (D) and glutamic acid (E) compared to phosphoserine. Image taken from (Zheng et al., 2019).

The mutations were created in the "sequence optimized" (CO) MmCpn60.1 gene. The double phosphomimetic mutant was created in all the three RBS spacer plasmids (i.e., 6, 8 and 10; see chapter 6, section 6.3.6). It was hypothesized that if phosphorylation at these residues is required for oligomer formation of MmCpn60.1, the double phophomimetic mutant of MmCpn60.1-CO would be able to perform the chaperonin function and sustain cell growth in the strain having repressed GroEL e.g., *E. coli* MGM100.

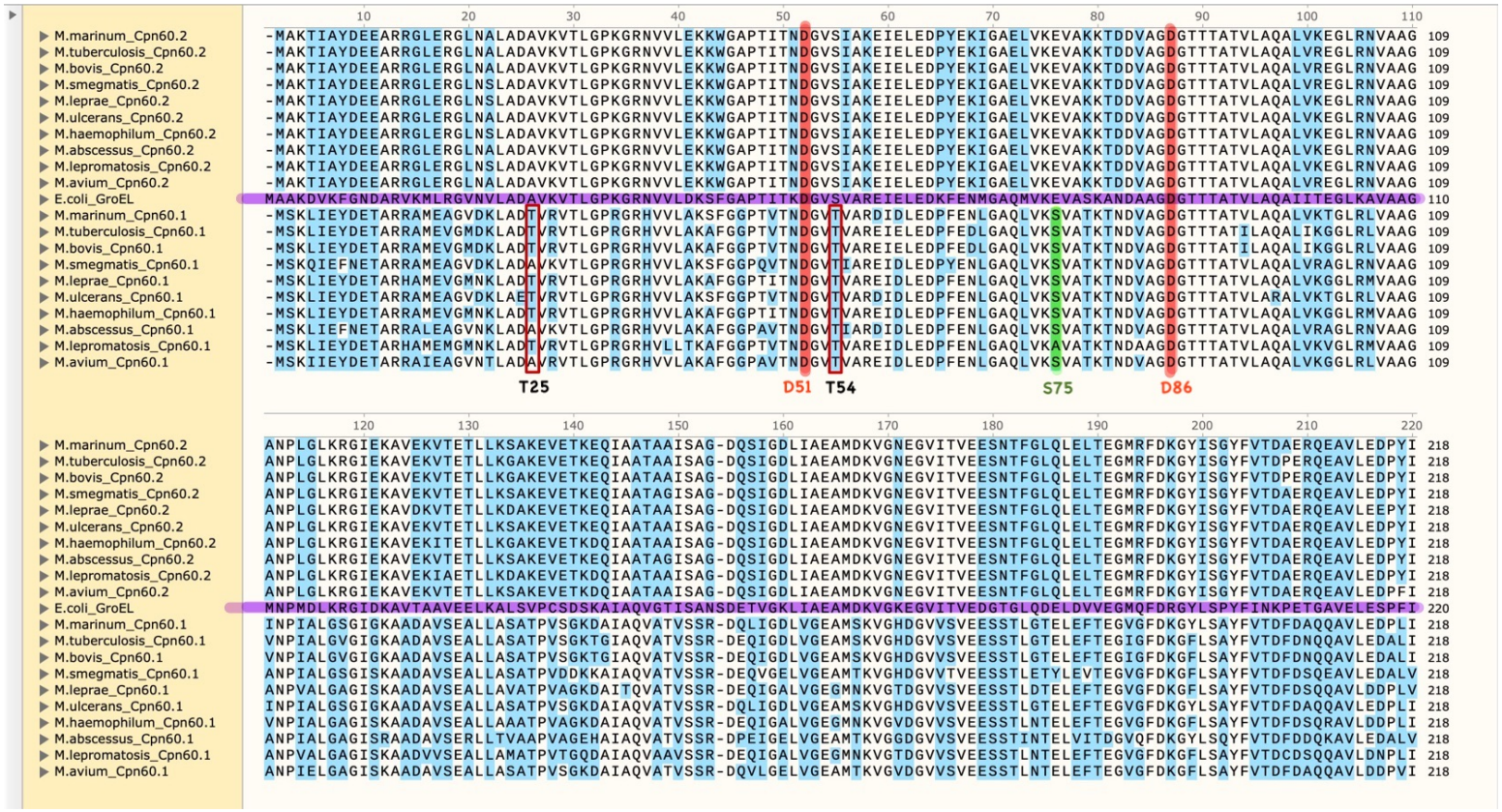


Figure 7.3 1-Multiple sequence alignment of Cpn60.2 and Cpn60.1 of ten species of mycobacteria.
For comparison *E. coli* GroEL (highlighted in purple) is in the middle. First 220 residues of GroEL and corresponding 218 residues of Cpn60.2s and Cpn60.1s are shown; remaining residues are shown in 7.4 & 7.5. The alignment was done using default parameters of Muscle alignment in Snapgene multiple sequence alignment tool. Serine (S) at position 75 of Cpn60.1 of all the mycobacteria highlighted in green; only *M. lepromatosis* Cpn60.1 have alanine (A) instead of serine at this position. At the corresponding position, GroEL and Cpn60.2 have a highly conserved glutamic acid (E). Residues not matching with the GroEL are highlighted in blue.

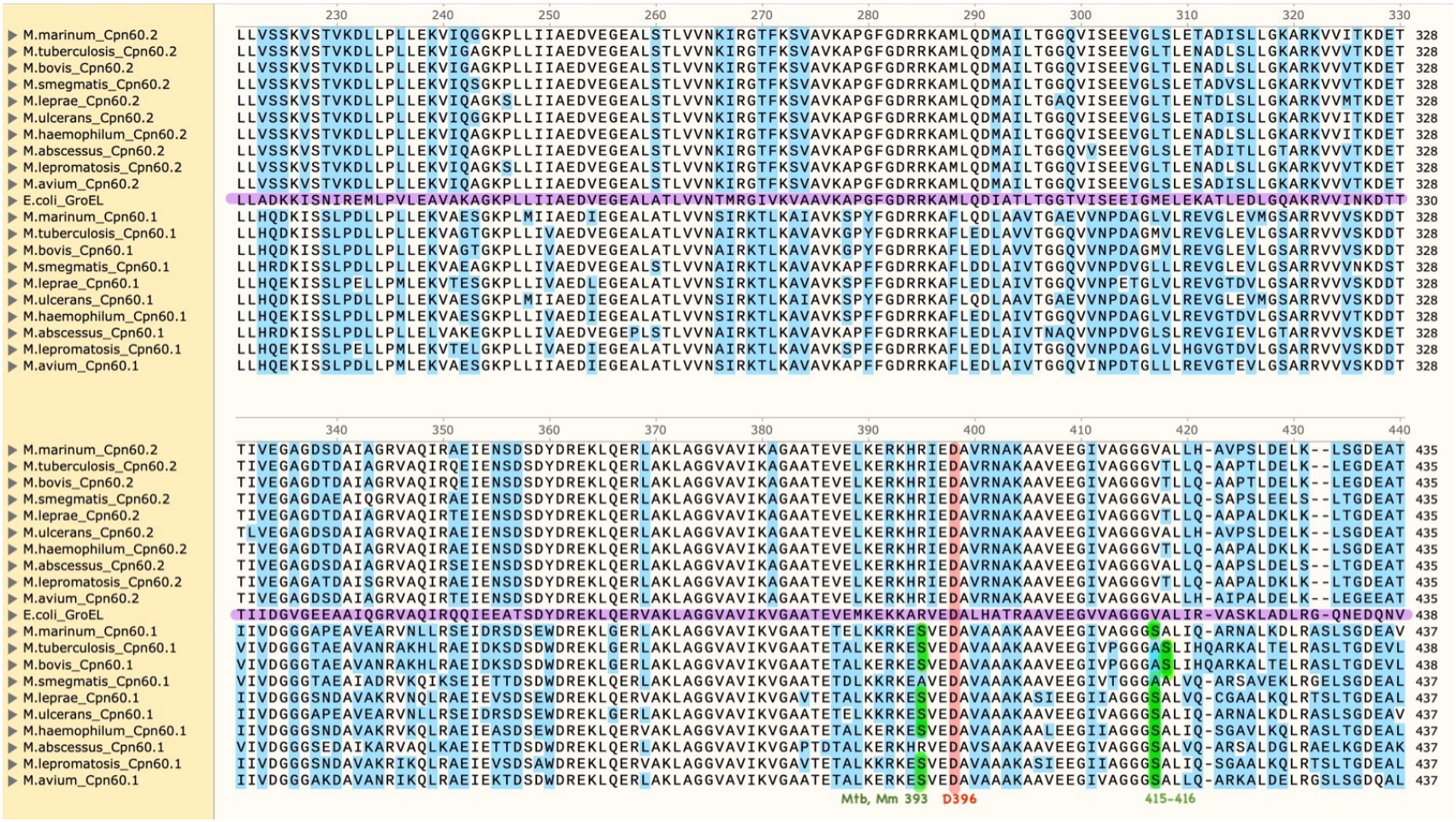


Figure 7.4 2-Multiple sequence alignment of Cpn60.2 and Cpn60.1 of ten species of mycobacteria.
Multiple sequence alignment (continued from 7.3, residues 221 to 438 of GroEL and the corresponding residues of Cpn60.2s and Cpn60.1s are shown. Serine (S) residues at position 393 of Cpn60.1 of 8 out of 10 mycobacterial species are marked in green. GroEL and Cpn60.2 have a highly conserved arginine (R). Serine at 416 position of MtbCpn60.1 is also marked in green; MmCpn60.1 has alanine at 416 position but contains serine at 415 position.

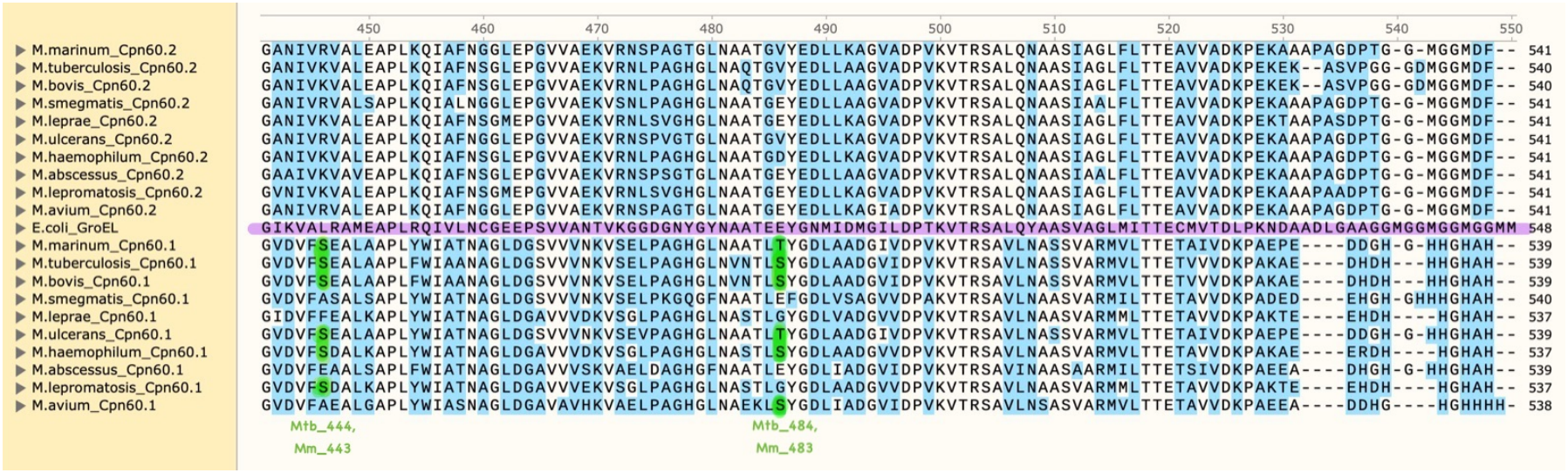


Figure 7.5 3-Multiple sequence alignment of Cpn60.2 and Cpn60.1 of ten species of mycobacteria.
Multiple sequence alignment (continued from 7.3 & 7.4), residues 439 to 548 of GroEL and the corresponding residues of Cpn60.2s and Cpn60.1s are shown. At position 444 MtbCpn60.1 contains serine whereas the corresponding serine in MmCpn60.1 is 443 position (marked in green) and at position 484 MtbCpn60.1 has serine residue whereas the corresponding position in MmCpn60.1 is threonine 483 (marked in green).

7.3 Results

7.3.1 Serine/threonine protein kinases of *M. mairnum* 1218R

The annotated genome sequence of 1218R was downloaded from NCBI (accession no. NZ_CP025779.1) and searched using the Artemis genome browser for open reading frame (ORFs) encoding serine/threonine protein kinases (STPKs). 18 potential STPKs were identified from the entire genome. The amino acid sequences of the ORFs encoding 18 STPKs of 1218R and 11 STPKs of H37Rv were collected and a multiple sequence alignment was performed in DNAstar Megalign pro software using default parameters of MUSCLE algorithm. The pairwise % similarity and % identity was determined based on the alignment (Figure 7.6). High sequence similarity (%) and identity (%) was observed between many enzymes.

The protein sequence of 1218R ORF MM1218R_RS13070 was found to have 94.85% similarity and 85.57% identity with PknF protein kinase of H37Rv (Figure 7.6, Table 7.1). The corresponding CDS for PknF gene of H37Rv was also visualised in the Artemis comparison tool and the same ORF MM1218R_RS13070 was found in correspondence with PknF (Figure 7.7). Among other ORFs having high sequence similarity with H37Rv STPKs are MM1218R_RS00095, MM1218R_RS22980, MM1218R_RS03345, and MM1218R_RS20905 which were found to have 94.85%, 96.91%, 97.94%, and 96.91% sequence similarity and 91.75%, 89.69%, 91.75%, and 92.78% sequence identity with PknA, PknD, PknG, and PknH of H37Rv respectively.

More than one ORF showed high levels of sequence similarity and identity with a single enzyme of H37Rv. For example, six ORFs showed more than 85% sequence similarity and more than 73% sequence identity with PknH (coded in green colour in table 7.1) and five ORFs were found to have more than 69% sequence similarity and more than 60% sequence identity with PknF of H37Rv (coded in red colour in Table 7.1). Some ORFs were also observed to have high similarity with more than one STPKs of H37Rv.

Both species of mycobacteria, H37Rv and 1218R, possess eukaryotic like serine/threonine protein kinases and many of them have high levels of homology. The data shows that the STPKs of both the organisms are highly similar which is consistent with the hypothesis that they might perform similar functions and phosphorylate the

Cpn60.1 at corresponding residues. In the following sections the creation of the phosphomimetic mutants of MmCpn60.1 and the complementation analysis will be discussed.

7.3.2 Single and double phosphomimetic mutations at S75 and S393 positions of Mmcpn60.1-CO

In the previous chapter (section 6.3.6), RBS spacer region optimization of the plasmid pTrc-Mm*cpn10*-Mm*cpn60.1*-CO containing the "sequence optimized" *cpn60.1* and *cpn10* of *M. marinum* was discussed. S75D and S393D mutations were created in pTrc-Mm*cpn10*-Mm*cpn60.1*-CO separately using standard site directed mutagenesis (SDM) protocol. The double phosphomimetic mutant of Mm*cpn60.1* (S75D+S393D) was created in all the three RBS distance optimized plasmids i.e., pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-6RBS, pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-8RBS, and pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-10RBS. The primers used for the SDM are given in chapter 2, section 2.3.

7.3.2.1 Complementation assay

The complementation assay was performed in *E. coli* MGM100 harbouring above plasmids. For comparison of the results MtbCpn60.1 W.T and double phosphomimetic mutant of MtbCpn60.1 were also included in the analysis. All strains grew well on 0.2% arabinose plate, as expected, because of the expression of chromosomal copy of *groESL* induced by arabinose in the media (Figure 7.8 A). When chromosomal copy of *groESL* was tightly repressed in the presence of glucose and in the absence of lactose or IPTG to induce the plasmid copy of the genes, no strain was expected to sustain growth. But it was observed previously that due to leaky expression of *trc* promoter the strains having strong chaperonin genes on pTrc plasmids were able to sustain growth (section 6.3.2) in these conditions. Two strains expressing *groESL* and Mtb*cpn10*-Mtb*cpn60.1*-DPM (S75D+S393D) at basal levels from pTrc plasmid showed full growth on 0.2% glucose plate (Figure 7.8 B).

Distance %Identity / %Similar Uncorrected Pairwise Distance

		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
A	Rv0015c_PknA		43.30	31.96	29.90	26.80	23.71	26.80	28.87	32.99	37.11	37.11	91.75	23.71	14.43	30.93	27.84	30.93	30.93	36.08	30.93	29.90	30.93	30.93	27.84	28.87	27.84	26.80	35.05	29.90
B	Rv0014c_PknB	62.89		43.30	41.24	42.27	26.80	42.27	32.99	51.55	38.14	46.39	45.36	24.74	18.56	37.11	41.24	41.24	40.21	37.11	40.21	43.30	43.30	46.39	41.24	38.14	39.18	36.08	46.39	38.14
C	Rv0931c_PknD	54.64	63.92		68.04	38.14	24.74	77.32	39.18	47.42	32.99	30.93	30.93	25.77	23.71	39.18	67.01	43.30	71.13	35.05	69.07	36.08	40.21	60.82	76.29	70.10	78.35	41.24	89.69	37.11
D	Rv1743_PknE	48.45	61.86	78.35		39.18	24.74	77.32	36.08	40.21	31.96	34.02	29.90	24.74	23.71	36.08	71.13	42.27	73.20	29.90	82.47	38.14	38.14	59.79	75.26	72.16	76.29	39.18	70.10	36.08
E	Rv1746_PknF	47.42	53.61	49.48	52.58		21.65	39.18	52.58	44.33	35.05	35.05	27.84	21.65	16.49	53.61	34.02	64.95	39.18	36.08	41.24	85.57	65.98	41.24	40.21	38.14	40.21	70.10	40.21	60.82
F	Rv0410c_PknG	41.24	45.36	41.24	41.24	37.11		23.71	29.90	26.80	26.80	27.84	22.68	91.75	16.49	29.90	23.71	23.71	28.87	26.80	27.84	22.68	22.68	23.71	23.71	22.68	24.74	22.68	24.74	21.65
G	Rv1266c_PknH	46.39	62.89	85.57	84.54	51.55	40.21		37.11	44.33	31.96	34.02	26.80	26.80	21.65	32.99	77.32	37.11	73.20	29.90	79.38	35.05	36.08	57.73	87.63	82.47	92.78	38.14	75.26	39.18
H	Rv2914c_PknI	44.33	50.52	51.55	50.52	65.98	45.36	51.55		41.24	34.02	28.87	28.87	30.93	21.65	65.98	37.11	52.58	36.08	34.02	40.21	53.61	55.67	36.08	36.08	35.05	35.05	58.76	40.21	49.48
I	Rv2088_PknJ	56.70	64.95	59.79	57.73	56.70	49.48	58.76	57.73		37.11	36.08	34.02	26.80	22.68	40.21	48.45	48.45	42.27	40.21	41.24	43.30	48.45	44.33	48.45	45.36	43.30	47.42	48.45	45.36
J	Rv3080c_PknK	55.67	54.64	47.42	44.33	45.36	42.27	42.27	47.42	51.55		29.90	35.05	26.80	16.49	40.21	31.96	37.11	34.02	81.44	34.02	35.05	37.11	31.96	32.99	35.05	31.96	31.96	32.99	32.99
K	Rv2176_PknL	52.58	61.86	53.61	55.67	50.52	46.39	54.64	49.48	54.64	48.45		37.11	29.90	14.43	30.93	36.08	38.14	35.05	31.96	38.14	35.05	35.05	39.18	31.96	30.93	31.96	34.02	31.96	40.21
L	MM1218R_RS00095	94.85	63.92	53.61	48.45	45.36	40.21	46.39	44.33	55.67	52.58	51.55		22.68	14.43	28.87	27.84	27.84	30.93	34.02	29.90	29.90	31.96	29.90	27.84	27.84	28.87	25.77	35.05	28.87
M	MM1218R_RS03345	41.24	46.39	41.24	41.24	37.11	97.94	42.27	46.39	48.45	43.30	48.45	40.21		16.49	28.87	26.80	21.65	28.87	27.84	28.87	23.71	22.68	21.65	27.84	24.74	28.87	21.65	23.71	24.74
N	MM1218R_RS03390	27.84	36.08	37.11	44.33	30.93	29.90	41.24	34.02	34.02	27.84	30.93	25.77	28.87		21.65	24.74	17.53	21.65	17.53	23.71	19.59	18.56	29.90	23.71	19.59	20.62	19.59	26.80	17.53
O	MM1218R_RS09015	45.36	49.48	49.48	50.52	69.07	43.30	48.45	74.23	55.67	46.39	47.42	42.27	44.33	34.02		32.99	52.58	40.21	37.11	36.08	54.64	55.67	35.05	31.96	34.02	30.93	57.73	36.08	48.45
P	MM1218R_RS09940	48.45	62.89	76.29	82.47	53.61	41.24	87.63	52.58	59.79	44.33	57.73	47.42	43.30	44.33	51.55		36.08	67.01	32.99	71.13	30.93	36.08	56.70	72.16	69.07	75.26	38.14	67.01	32.99
Q	MM1218R_RS12030	51.55	56.70	53.61	52.58	71.13	43.30	47.42	70.10	64.95	54.64	55.67	48.45	43.30	35.05	69.07	53.61		41.24	40.21	44.33	62.89	65.98	46.39	41.24	41.24	39.18	64.95	45.36	57.73
R	MM1218R_RS12215	50.52	58.76	82.47	81.44	52.58	45.36	85.57	49.48	58.76	44.33	55.67	48.45	45.36	38.14	48.45	79.38	52.58		34.02	76.29	37.11	38.14	59.79	70.10	70.10	74.23	39.18	72.16	40.21
S	MM1218R_RS12915	55.67	53.61	46.39	45.36	47.42	43.30	41.24	49.48	53.61	88.66	49.48	51.55	43.30	27.84	51.55	43.30	58.76	45.36		31.96	37.11	37.11	36.08	30.93	32.99	30.93	31.96	35.05	32.99
T	MM1218R_RS12945	48.45	61.86	80.41	85.57	51.55	43.30	86.60	54.64	56.70	43.30	57.73	47.42	44.33	41.24	51.55	80.41	53.61	84.54	44.33		42.27	40.21	64.95	77.32	73.20	80.41	42.27	72.16	42.27
U	MM1218R_RS13070	48.45	55.67	51.55	50.52	94.85	39.18	48.45	65.98	59.79	46.39	52.58	46.39	39.18	34.02	70.10	52.58	75.26	49.48	49.48	51.55		67.01	40.21	35.05	36.08	36.08	65.98	40.21	62.89
V	MM1218R_RS14825	48.45	55.67	50.52	50.52	74.23	40.21	48.45	70.10	61.86	50.52	50.52	49.48	41.24	32.99	70.10	51.55	79.38	49.48	52.58	51.55	76.29		39.18	36.08	36.08	35.05	63.92	45.36	59.79
W	MM1218R_RS17255	50.52	64.95	73.20	70.10	51.55	38.14	71.13	51.55	60.82	44.33	59.79	49.48	39.18	41.24	51.55	73.20	57.73	71.13	49.48	74.23	53.61	54.64		56.70	58.76	58.76	38.14	63.92	40.21
X	MM1218R_RS20820	48.45	62.89	83.51	83.51	52.58	42.27	95.88	52.58	61.86	44.33	54.64	49.48	43.30	40.21	48.45	86.60	49.48	86.60	43.30	86.60	49.48	50.52	69.07		80.41	86.60	39.18	74.23	37.11
Y	MM1218R_RS20890	49.48	59.79	82.47	81.44	48.45	40.21	91.75	48.45	58.76	47.42	54.64	47.42	42.27	34.02	49.48	85.57	49.48	84.54	45.36	83.51	46.39	47.42	70.10	89.69		81.44	37.11	70.10	38.14
Z	MM1218R_RS20905	47.42	61.86	86.60	82.47	50.52	41.24	96.91	51.55	57.73	41.24	54.64	48.45	43.30	39.18	48.45	86.60	47.42	85.57	41.24	86.60	47.42	46.39	72.16	94.85	91.75		36.08	75.26	39.18
AA	MM1218R_RS20920	44.33	47.42	50.52	51.55	79.38	38.14	48.45	69.07	59.79	45.36	51.55	42.27	38.14	34.02	69.07	52.58	75.26	50.52	46.39	52.58	79.38	75.26	52.58	50.52	46.39	48.45		45.36	57.73
AB	MM1218R_RS22980	53.61	65.98	96.91	82.47	51.55	40.21	87.63	53.61	63.92	46.39	53.61	52.58	40.21	39.18	48.45	79.38	54.64	82.47	45.36	82.47	53.61	53.61	74.23	84.54	82.47	86.60	54.64		39.18
AC	MM1218R_RS24715	43.30	52.58	52.58	52.58	69.07	42.27	54.64	60.82	61.86	47.42	54.64	43.30	44.33	35.05	59.79	52.58	71.13	55.67	48.45	56.70	72.16	72.16	52.58	54.64	53.61	55.67	68.04	52.58	

Figure 7.6 serine/threonine protein kinases of *M. tuberculosis* H37Rv and *M. marinum* 1218R genome.
Pairwise % identity (upper right; green) and % similarity (lower left; orange) of all the serine/threonine protein kinases of *M. tuberculosis* H37Rv and *M. marinum* 1218R genome. The highest values of identity and similarity for one enzyme are boxed in red. The intensity of the green or orange colour is directly proportional to the % identity/similarity.

Sr. No.	*M. tuberculosis* Ser/Thr protein kinases			Sr. No.	*M. marinum* Ser/Thr protein kinases				
	Name	ORF designation	M.W		ORF designation	M.W	Potential Mtb homologue	% Similarity	% Identity
1	PknA	Rv0015c	45597.59	1	MM1218R_RS00095	48105.48	PknA	94.85	91.75
2	PknB	Rv0014c	66510.48						
3	PknD	Rv0931c	69513.75	2	MM1218R_RS22980	73387.16	PknD	96.91	89.69
4	PknE	Rv1743	60512.617						
5	PknF	Rv1746	50668.54	3	MM1218R_RS13070	50301.184	PknF	94.85	85.57
6	PknG	Rv0410c	81578.29	4	MM1218R_RS03345	83321.25	PknG	97.94	91.75
7	PknH	Rv1266c	66754.22	5	MM1218R_RS20905	70223.92	PknH	96.91	92.78
8	PknI	Rv2914c	61805.25	6	MM1218R_RS09015	62830.625	PknI	74.23	65.98
9	PknJ	Rv2088	61563.836						
10	PknK	Rv3080c	119419.96	7	MM1218R_RS12915	120023.4	Pknk	88.66	81.44
11	PknL	Rv2176	42802.16						
				8	MM1218R_RS03390	18251.133			
				9	MM1218R_RS12030	53364.35	PknF	71.13	64.95
				10	MM1218R_RS12215	59865.6	PknH	85.57	73.20
				11	MM1218R_RS12945	63893.3	PknH	86.60	79.38
				12	MM1218R_RS14825	69731.3	PknF/PknI	74.23/70.10	65.98/55.67
				13	MM1218R_RS09940	64215.39	PknH	87.63	77.32
				14	MM1218R_RS20890	65870.19	PknH/PknD	91.75/82.47	82.47/70.10
				15	MM1218R_RS20820	63579.035	PknH	95.88	87.63
				16	MM1218R_RS20920	52507.836	PknF	79.38	70.10
				17	MM1218R_RS17255	66182.71	PknD	73.20	60.82
				18	MM1218R_RS24715	53623.926	PknF	69.07	60.82

Table 7.1 Serine/threonine protein kinases of *M. tuberculosis* H37Rv and *M. marinum* 1218R.
M.W is given in Daltons. For *M. marinum* STPKs the potential Mtb homologous enzymes are mentioned according to the sequence similarity (%) and identity (%). Potential homologs are colour coded.

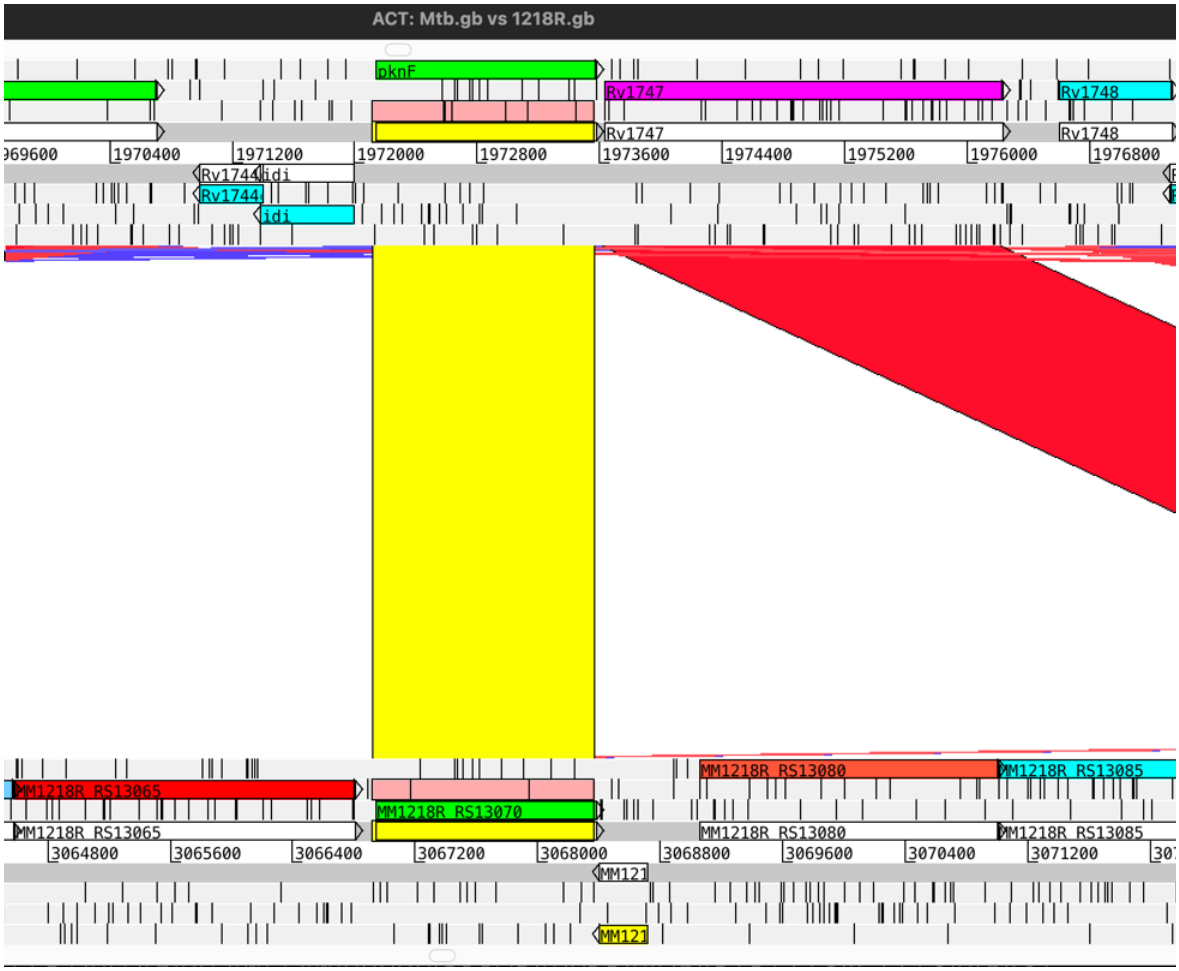


Figure 7.7 Visualization of PknF of H37Rv and the corresponding gene of 1218R in Artemis comparison tool.
Both the CDS are in + orientation on the chromosome and have high sequence identity and similarity.

On 0.1mM IPTG plate (in the absence of arabinose), apart from the two strains which showed growth on plate B, no other strain could sustain growth till the first 24 hours of incubation (Figure 7.8 C). Upon longer incubation up to 40 hours two strains harbouring plasmids pTrc-Mm*cpn10*-Mm*cpn60.1*-CO and pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-10RBS showed colonies up to 10^{-3} dilution (Figure 7.8 D) however, the effect is negligible and does not appear to require DPM. No single phosphomimetic mutant of MmCpn60.1 was able to perform the chaperone function in *E. coli* irrespective of the length of the incubation.

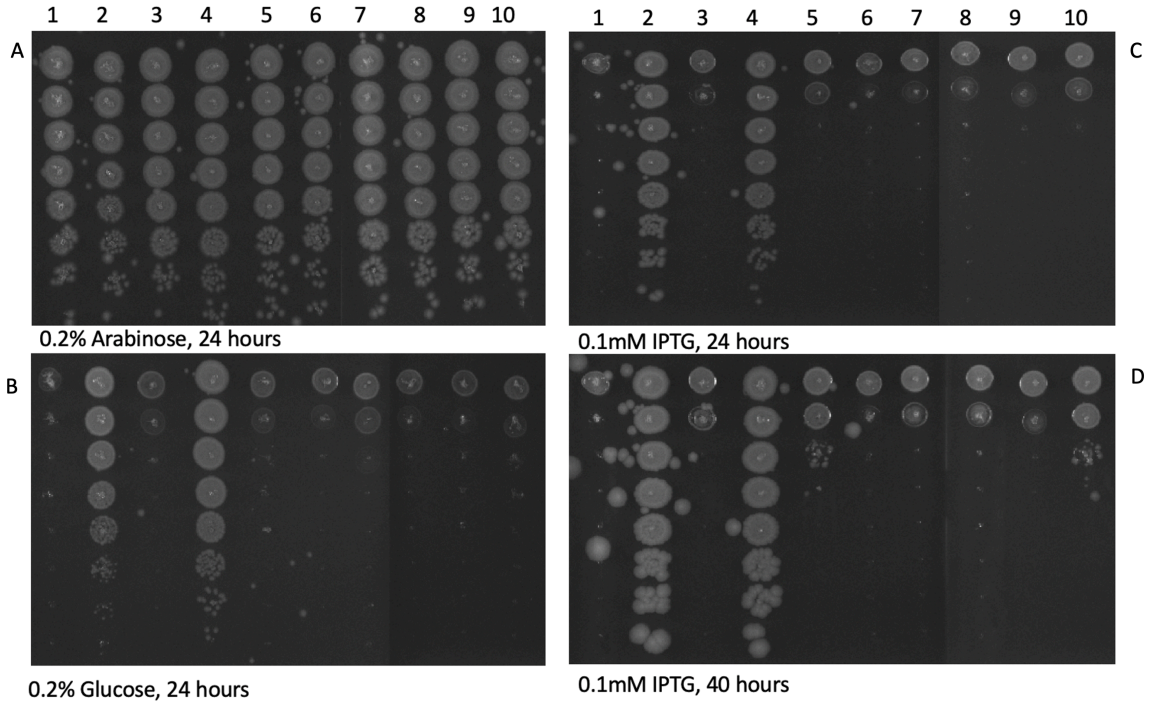


Figure 7.8 Complementation assay of single and double phosphomimetic mutants of MmCpn60.1-CO.

Complementation assay in MGM100 harbouring pTrc99A derivative plasmids containing single and double phosphomimetic mutants of Mm*cpn60.1*-CO. All plates were incubated at 30°C. kanamycin (selection marker for *E. coli* MGM100) and ampicillin (selection marker for the plasmid) were added in all plated. 1) pTrc99A, 2) pTrc-*groES-groEL*, 3) pTrc-Mtb*cpn10*-Mtb*cpn60.1*, 4) pTrc-Mtb*cpn10*-Mtb*cpn60.1*-DPM, 5) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO, 6) pTrc-Mm*cpn10*-Mm*cpn60.*1-CO-S75D, 7) pTrc-Mm*cpn10*-Mm*cpn60.*1-CO-S393D, 8) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-6RBS, 9) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-8RBS, and 10) pTrc-Mm*cpn10*-Mm*cpn60.*1-CO-DPM-10RBS.

7.3.2.2 Expression analysis

On SDS-PAGE gel, pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-8RBS showed detectable levels of expression (Figure 7.9, lane 8).

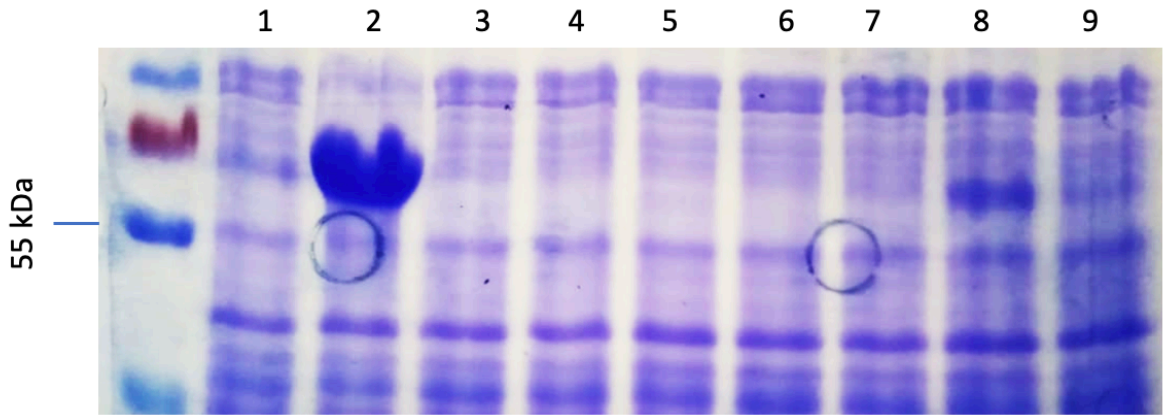


Figure 7.9 Overexpression of double RBS optimized phosphomimetic mutants of *M. marinum* Cpn60.1-CO.
Overexpression of double phosphomimetic mutants of *M. marinum* Cpn60.1-CO on SDS-PAGE gel in Tab21 strain induced at 30°C for 4 hours with 1mMIPTG. (1) pTrc99A, (2) pTrc-*groES-groEL*, (3) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO, (4) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-S75D, (5) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-S393D, (6) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM, (7) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-6RBS, (8) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-8RBS, (9) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-10RBS.

7.3.3 Phosphomimetic mutations of serine and threonine residues in the equatorial domain (towards C-terminal) of MmCpn60.1

The complementation results of MmCpn60.1 DPM (S75D+S393D) were different than the results of MtbCpn60.1 DPM; the double phosphomimetic mutant (S75D+S393D) of MtbCpn60.1 was able to sustain cell growth in combination with MtbCpn10 in the tight repression of chromosomal *groESL* (Figure 7.8 B) whereas that of MmCpn60.1 could not sustain cell growth even after 40 hours of incubation (Figure 7.8 D). One possible reason could be that MmCpn60.1 might require phosphorylation at some additional residues to form higher order oligomers. To check this possibility, some additional serine and threonine sites at the inter-subunit interface were chosen to modify in the DPM of Mm*cpn60.1*. The sites chosen were S415, S444, T481 and T483 (Figure 7.3, Figure 7.4, & Figure 7.5). It was hypothesized that one of the triple phosphomimetic mutant might complement the loss of GroEL in MGM100. In the double phosphomimetic mutant of pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM (S75D+S393D) with

8 bps spacer RBS phosphomimetic mutations (because of detectable expression levels on SDS-PAGE) were constructed at four sites S415D, S443D, T481D, and T483D. The mutants of Mm*cpn60.1*-CO-DPM-8RBS with additional mutations at S415D, 8RBS, S443D, T483D, T481D, and T481D+T483D in pTrc plasmid with Mm*cpn10*-CO were tested for complementation in MGM100.

The results of the complementation assay are shown in Figure 7.8. All strains grew well on arabinose plate. On glucose plate only the strain expressing groESL from the leaky *trc* promoter could sustain growth (positive control). On IPTG and lactose plates, none of the triple phosphomimetic MmCpn60.1 mutants could complement the loss of GroEL even after 40 hours of incubation at 30°C. The quadruple also failed to complementation the loss of GroEL in *E. coli* (Figure 7.10).

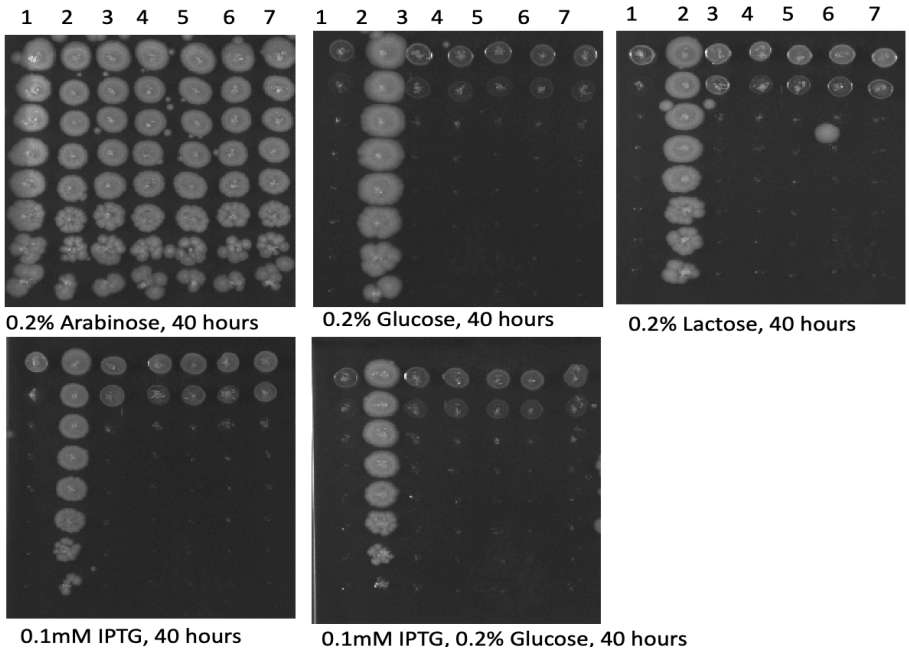


Figure 7.10 Complementation assay of various double phosphomimetic mutants of MmCpn60.1-CO.
Complementation assay in MGM100 harbouring pTrc99A derivative plasmids containing various phosphomimetic mutants of Mm*cpn60.1*-CO-8RBS. All plates were incubated at 30°C for 40 hours. kanamycin and ampicillin were added in all plated. 1) pTrc99A, 2) pTrc-*groES-groEL*, 3) pTrc-Mtb*cpn10*-Mtb*cpn60.1*-CODPM-S415D-8RBS, 4) pTrc-Mtb*cpn10*-Mtb*cpn60.1*-CO-DPM-S443D-8RBS, 5) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-T483D-8RBS, 6) pTrc-Mm*cpn10*-Mm*cpn60.*1-CO-DPM-T481D-8RBS, 7) pTrc-Mm*cpn10*-Mm*cpn60.*1-CO-DPM-T481D+T483D-8RBS.

7.3.4 Complementation analysis of WT-CO and DPM of Mm*cpn60.1* in combination with *E. coli groES*.

The results of the complementation assay with Mmcpn60.1-CO (W.T) and Mmcpn60.1-CO-DPM were unexpected. None of the *M. marinum* Cpn60.1 could complement the loss of GroEL in *E. coli* in combination with its cognate co-chaperonin MmCpn10. One possible reason could be that the co-chaperonin MmCpn10 could not work optimally in *E. coli*. To investigate if changing the co-chaperonin had any effect on the complementation ability of the "sequence optimized" and DPM of *M. marinum* Cpn60.1, the Mmcpn10 gene was swapped with *E. coli groES* in three plasmids pTrc-Mm*cpn10*-Mm*cpn60.1*-CO, pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM and pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-8RBS. Among RBS distance optimised plasmids, 8bps spacer plasmid was used because the 8RBS plasmid showed detectable levels of expression on SDS-PAGE gel and western blot (Figure 7.9).

To prepare pTrc99A plasmid derivatives pTrc-*groES*-Mm*cpn60.1*-CO, pTrc-*groES*-Mm*cpn60.1*-CO-DPM and pTrc-*groES*-Mm*cpn60.1*-CO-DPM-8RBS, a *groES* containing pTrc plasmid (section 6.3.1.1) pTrc-*groES*-Mm*cpn60.2* was used. In this plasmid the Mm*cpn60.2* was swapped with Mm*cpn60.1*-CO, Mm*cpn60.1*-CO-DPM, and Mm*cpn60.1*-CO-DPM-8RBS between BglII and HindIII sites. The protocol followed for the cloning steps such as restriction digestion, ligation and transformation are given in chapter 2, section 2.6.2.

The complementation assay in MGM100 harbouring the above plasmids was performed. Strains harbouring pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM and pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-8RBS were also included in the analysis for comparison of results. The results of the complementation assay are shown in figure 7.9. No strain could grow in first 24 hours of incubation at 30°C on IPTG plate except the positive control (plasmid copy of *groES-groEL* expressing strain). A longer incubation up to 40 hours showed considerable growth (up to 10^{-5} dilution) of all the three strains expressing plasmid copy of *groES* in combination with Mm*cpn60.1* irrespective of the phosphorylation and RBS distance optimization status (Figure 7.11). It was unexpected that MmCpn60.1-CO would show complementation in combination with GroES. MmCpn60.1-CO-DPM did not show complementation in combination with MmCpn10 even after 40 hours of incubation as found in previous assay (Figure 7.8).

The reason for complementation showed by MmCpn60.1-CO and DPM in combination with GroES but not MmCpn10 was not known. One reason could be the expression levels. MmCpn60.1 might express at higher levels when co-expressed with GroES rather than MmCpn10. To test his hypothesis, western blotting was performed.

If MmCpn60.1-CO, MmCpn60.1-CO-DPM, and MmCpn60.1-CO-DPM-8RBS in combination with GroES oligomerize and perform a functional chaperonin activity then the ATPase mutations in these chaperonins should abolish the chaperonin ability as was seen in the experiments of Dr Santosh Kumar where MtbCpn60.1-DPM lost the complementation ability due to ATPase mutations at 51, 86 and 396 residues (D51A, D86A, and D396A) (Figure 7.2). The next two sections describe the western blotting and ATPase mutations experiments done to test the above two hypotheses.

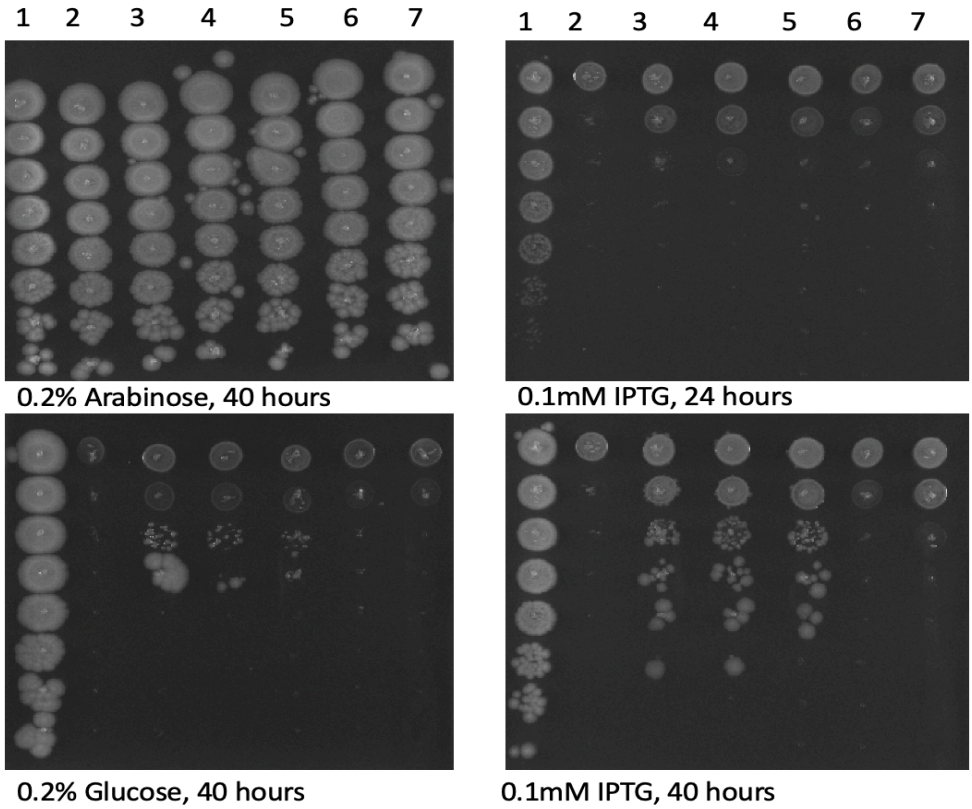


Figure 7.11 Complementation analysis of WT-CO and DPM of Mm*cpn60.1* in combination with *E. coli groES*.
Complementation Assay in MGM100. All plates contained kanamycin and incubated at 30°C. 1) pTrc-*groES-groEL*, 2) pTrc99A, 3) pTrc-*groES*-Mm*cpn60.*1-CO-DPM-8RBS, 4) pTrc-*groES*-Mm*cpn60.1*-CO-DPM, 5) pTrc-*groES*-Mm*cpn60.1*-CO, 6) pTrc-Mm*cpn10*-Mm*cpn60.1*-DPM, and 7) pTrc-Mm*cpn10*-Mm*cpn60.1*-DPM-8RBS.

7.3.5 Western blotting experiment to compare the expression levels of MmCpn60.1-CO (W.T, DPM, and DpM-8RBS) expressed in combination with GroES and MmCpn10

To test if the expression levels of MmCpn60.1-CO W.T and DPM were higher in case of co-expression with GroES than with MmCpn10, western blotting was performed. MGM100 strains harbouring pTrc plasmids derivates containing Mm*cpn60.1*-CO W.T, DPM, and DPM-8RBS in combination with *groES* or Mm*cpn10* were used to prepare the sample for SDS-PAGE gel.

To overexpress the plasmid copy of chaperonins, IPTG was used to induce the secondary cultures for 6 hours. The cell lysates of all the samples were loaded on SDS-PAGE gel. The proteins were transferred from SDS-PAGE to Polyvinylidene difluoride (PVDF) membrane using electroblotting. The PVDP membrane was treated with primary and secondary anti-rabbit IgG antibody specific to MmCpn60.1. The detailed procedure of SDS-PAGE and western blotting is given in chapter 2, section 2.7. The western blot is shown in Figure 7.12.

Lane 1 contained GroEL as a negative control. Lanes 2 and 5 contained MmCpn60.1-CO expressed from pTrc plasmids in combination with MmCpn10 and GroES respectively. Lanes 3 and 6 contained MmCpn60.1-CO-DPM (with natural RBS spacer) expressed from pTrc plasmid in combination with MmCpn10 and GroES respectively and lanes 4 and 7 contained MmCpn60.1-CO-DPM-8RBS expressed in combination with MmCpn10 and GroES respectively.

No noticeable difference was observed in the expression levels of any of the chaperonins either expressed in combination with MmCpn10 or GroES. However, the expression of MmCpn60.1-CO-DPM-8RBS was found to be the highest compared to the W.T and DPM with normal RBS spacer and the expression of DPM was observed to be greater than W.T.

The slight difference in the molecular weight of W.T and DPM could be because aspartic acid (133Da) has higher molecular weight than serine (105Da). Western blot also showed a small amount of degraded MmCpn60.1 in all the samples. As the antibody specifically binds to the C-terminal part of the protein, the degradation might had been done from the N-terminal part.

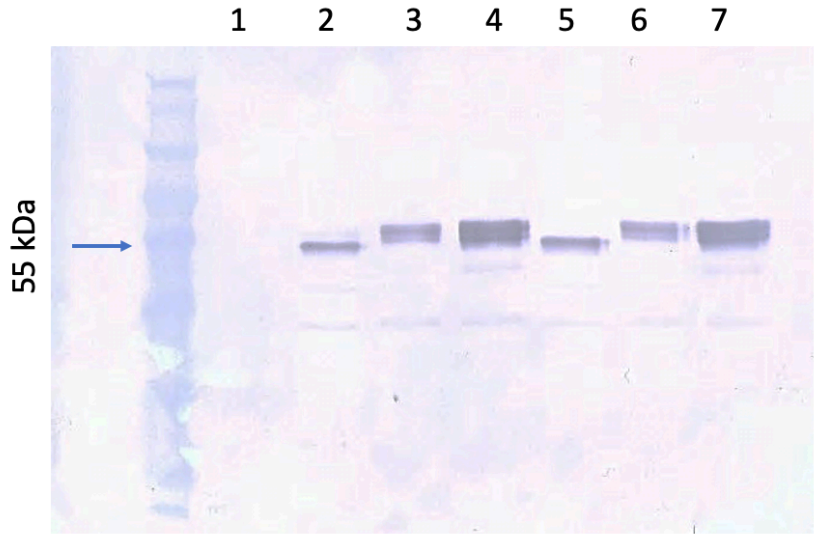


Figure 7.12 Western blot of MmCpn60.1-CO-DPM expressed from plasmids in MGM100 strain. 1) pTrc-*groESL*, 2) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO, 3) pTrc-Mm*cpn10*-Mm*cpn60.1*-DPM, 4) pTrc-Mm*cpn10*-Mm*cpn60.1*-DPM-8RBS, 5) pTrc-*groES*-Mm*cpn60.1*-CO, 6) pTrc-*groES*-Mm*cpn60.1*-DPM, 7) pTrc-*groES*-Mm*cpn60.1*-DPM-8RBS.

7.3.6 Effect of ATPase mutations on the complementation ability of MmCpn60.1-CO proteins (W.T, DPM, DPM-8RBS) in combination with GroES.

To test if the partial complementation shown by pTrc-*groES*-Mm*cpn60.1*-CO, pTrc-*groES*-Mm*cpn60.1*-CO-DPM, and pTrc-*groES*-Mm*cpn60.1*-CO-DPM-8RBS in MGM100 upon 40 hours of incubation at 30°C (Figure 7.11) was due to the functional chaperonin activity of MmCpn60.1 protein expressed by these plasmids, various ATPase mutations were created in the Mmcpn60.1 gene present in all the three plasmids. For this purpose, aspartic acid (D) residues at positions 51, 86 and 396 were mutated to alanine (A) (Figure 7.3 & Figure 7.4). Table 7.2 shows different ATPase mutations created in the Mm*cpn60.1* gene present in the above three plasmids.

While creating ATPase mutations, one mutant (created by chance) was found to have frameshift mutation in Mm*cpn60.1* gene after 213 bps (71 residues). If the complementation observed in figure 7.9 was due to the complementation ability of MmCpn60.1 then the frameshift mutation would abolish this phenotype. This mutant was therefore also included in the complementation analysis. All the derivative plasmids containing ATPase mutations were transformed in MGM100 and used for complementation assay.

It was demonstrated in section 7.3.2.1 neither of the three MmCpn60.1-CO proteins (W.T, DPM, DPM-8RBS) showed complementation in combination with MmCpn10. But the same three proteins demonstrated complementation ability in combination with GroES (examined in section 7.3.4). It was important to check whether high level expression of GroES alone could be the cause of the apparent complementation. Therefore, a pTrc derivative plasmid obtained from the lab stock containing *groES* alone (pTrc-*groES*-SpeI) regulated by *trc* promoter was also included in the analysis.

Sr No.	Plasmid	Showed Complementation on IPTG plate after 40 hours of incubation	ATPase mutation done	Name given to plasmid after ATPase mutation made in Mm*cpn60.1*
1	pTrc-*groES*-Mm*cpn60.1*-CO	++	D51A	CO-D51A
			D396A	CO-D396A
			D51A+D86A	CO-D51A+D86A
			D51A+D86A+D396A	CO+D51A+D86A+D396A
2	pTrc-*groES*-Mm*cpn60.1*-CO-DPM	++	D86A	CO-DPM-D86A
			D51A	DO-DPM-D51A
3	pTrc-*groES*-Mm*cpn60.1*-CO-DPM-8RBS	++	D51A,	CO-CPM-8-D51A
			D86A	CO-DPM-8-D86A
			D51A+D86A	CO-DPM-8-D51A+D86A
			Frameshift mutation	CO-DPM-8-F

Table 7.2 ATPase mutations created in Mm*cpn60.1*-CO (W.T, DPM, DPM-8RBS).

The result of the complementation assay is shown in Figure 7.13. All the strains showed optimal growth in the presence of arabinose when chromosomal *groESL* was expressed (Figure 7.13 A, B). On glucose plate, up to first 24 hours of incubation only the strain expressing groESL from the pTrc plasmid due to leaky nature of trc promoter showed growth whereas upon longer incubation for 40 hours all other strains apart from pTrc99A (negative control) showed negligible growth up to the 2nd highest dilution (Figure 7.13 C, D).

On IPTG plate, strain harbouring GroES only plasmid (pTrc-*groES*-SpeI) showed growth up to 10^{-3} dilution whereas all other strains harbouring plasmids 4-9 & 10-16 showed growth up to 10^{-4} dilution (Figure 7.13 E, F). All strains could sustain partial growth irrespective of the status of the MmCpn60.1-CO expressed from the plasmid; the W.T, DPM, DPM-8RBS showed partial complementation as observed previously (Figure 7.11) but it was expected that the ATPase mutations would abolish the complementation ability. Unexpectedly, ATPase mutations showed the same level of growth as that of phosphomimetic mutations. Even the growth rate was not observed

to be different in case of frameshift mutation in MmCpn60.1-CO. Results of IPTG only and IPTG plus glucose plates were identical to each other (Figure 7.13 G, H). The result show that the maintenance of partial growth was probably not due to the complementation ability of MmCpn60.1 mutants but the excessive GroES expressed from the plasmid. A high level of expression of GroES alone is enough to see some complementation. The likely interpretation is that in MGM100, the first limiting factor for growth is lack of GroES. However, if we continue to grow these strains (e.g., in liquid medium) even in the presence of high GroES that lack of GroEL would eventually cause them to stop growing.

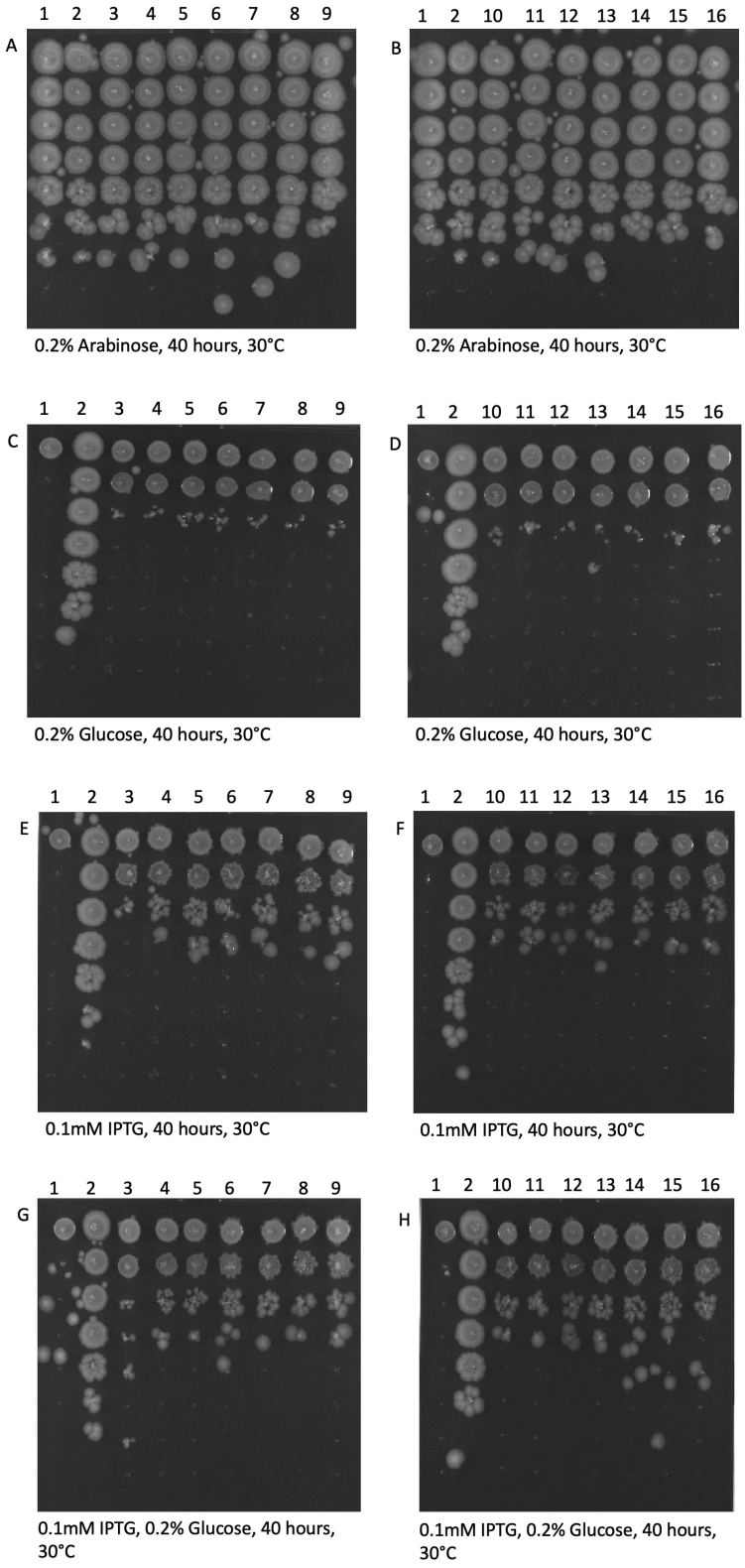


Figure 7.13 Complementation assay of ATPase mutants in MGM100.
Complementation assay in MGM100 harbouring plasmids 1-16. All plates contained kanamycin and ampicillin. 1) pTrc99A, 2) pTrc-*groESL*, 3) pTrc-*groES*-SpeI, 4) pTrc- pTrc-*groES*-Mm*cpn60.1*-CO, 5) pTrc-*groES*-Mm*cpn60.1*-CO-DPM, 6) pTrc-*groES*-Mm*cpn60.1*-CO-DPM-8RBS, 7) CO-D51A, 8) CO-D396A, 9) CO-D51A+D396A, 10) CO+D51A+D86A+D396A, 11) CO-DPM-D86A, 12) DO-DPM-D51A, 13) CO-CPM-8-D51A, 14) CO-DPM-8-D86A, 15) CO-DPM-8-D51A+D86A, and 16) CO-DPM-8-F.

7.3.7 Effect of swapping of C-termini between Mtb*cpn60.1*-DPM and Mm*cpn60.1*-DPM on complementation ability of phosphomimetic mutants

In the above sections it was demonstrated that double phosphomimetic mutant of MtbCpn60.1 in combination with its co-chaperonin MtbCpn10 showed functional chaperonin activity by complementing the loss of GroESL (tightly suppressed) in MGM100 whereas, the double phosphomimetic mutant of MmCpn60.1 could not sustain cell growth even after longer incubation of 40 hours. This suggested that lack of phosphorylation of MmCpn60.1 was probably not the reason why complementation was not seen. This made us consider alternative explanations for why MmCpn60.1 failed to complement in *E coli*.

It is possible that other structural differences between MTb and MmCpn60.1 might explain why the DPM of one of these proteins functions in *E. coli* and the other does not. Such differences might be revealed by aligning the two proteins and searching for regions of diversity in amino-acid sequence. We therefore aligned the amino acid sequences of both the chaperonins to find the regions of diversity and similarity. The sequence identity of both the proteins is 87.08% whereas the sequence similar is 94.01%. The most diverse areas were found to be present towards C-termini of the proteins (Figure 7.14). Due to high diversity in the C-termini it was hypothesized that the differences in the C-termini might account for the difference in the complementation data. Therefore, the C-termini of MtbCpn60.1-DPM and MmCpn60.1-CO-DPM-8RBS were swapped to investigate the importance of the C-terminal sequence. It was hypothesized that the MmCpn60.1-CO-DPM with Mtb C-terminal might be able to form functional oligomers and could complement the loss of GroEL in *E. coli*.

It was observed that the sequence of Mtb*cpn60.1* naturally contained a HindIII site 531 bps upstream the stop codon. A project student in our lab, Doaa Ghulam, inserted the HindIII site in Mm*cpn60.1* present in the plasmid pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-8RBS at the same position (this (luckily) did not cause any change in amino acid sequence.). Both the plasmids contained another HindIII site after *cpn60.1* gene. I sequenced the plasmid in which the HindIII site was inserted by Doaa Ghulam and confirmed that the sequence was correct and did not have any unwanted mutations. I

used the verified plasmid and pTrc-Mtbc*pn10*-Mtb*cpn60.1*-DPM to swap the HindIII fragments of Mtb*cpn60.1* and Mm*cpn60.1* (Figure 7.15). Both the plasmids were digested by HindIII, and the fragments and the plasmid backbones were purified by gel electrophoresis. The fragments were than ligated and the ligation reaction was transformed in high efficiency TOP10 cells.

The plasmid constructs pTrc-Mtb*cpn10*-Mtb*cpn60.1*_MmHindIII_DPM and pTrc-Mm*cpn10*-Mm*cpn60.1*_MtbHindIII_DPM were screened, purified, and verified by sequencing. The plasmids were transformed in MGM100, and complementation analysis was performed.

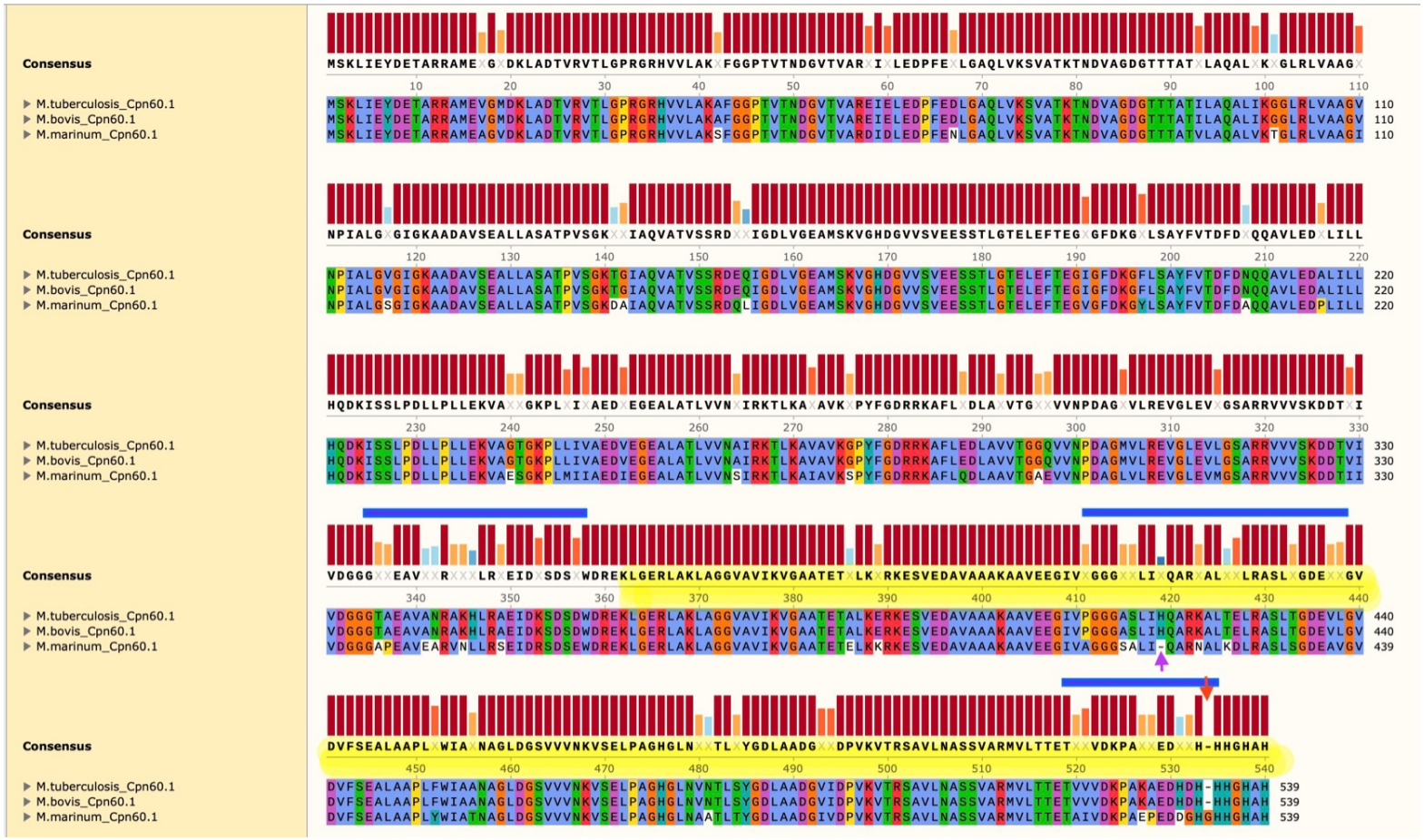


Figure 7.14 Sequence alignment of *M. tuberculosis* Cpn60.1, *M. bovis* Cpn60.1 and *M. marinum* Cpn60.1 proteins.
Alignment was done using default parameters of MUSCLE alignment. Most diverse regions are marked with blue blocks. Sequence was found to be more diverse at C-terminal than N-terminal region. Two gaps are marked with purple and red arrows. HindIII fragment swapped between MTb*cpn60.1*-DPM and Mm*cpn60.1*-CO-DPM is highlighted in yellow.

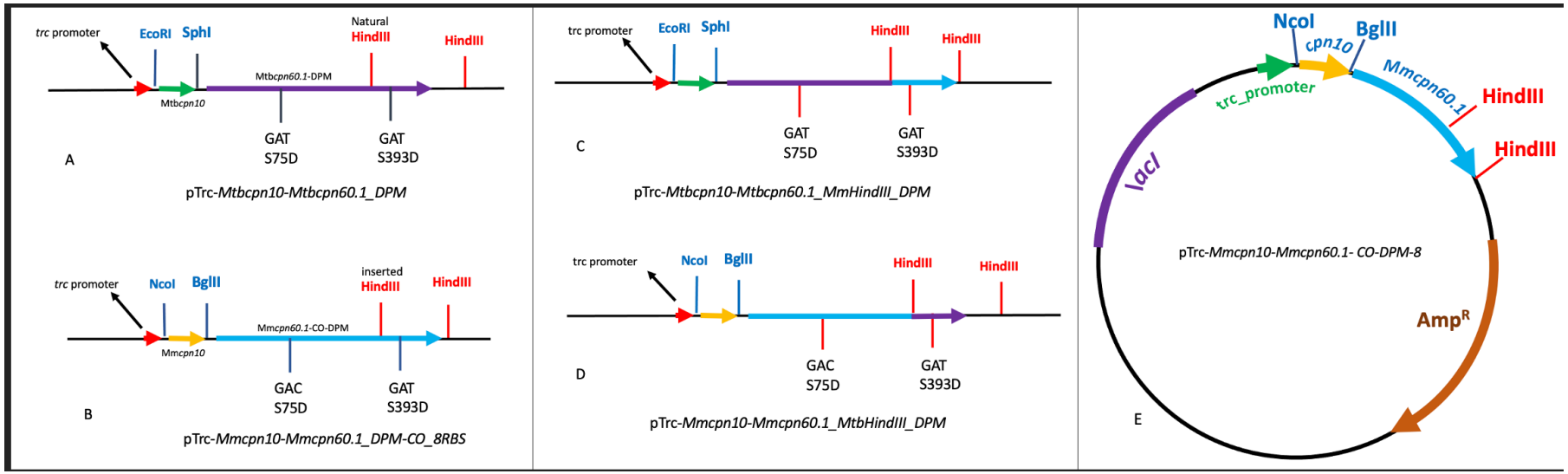


Figure 7.15 Schematic plan for swapping of HindIII fragment between Mtb*cpn60.1*-DPM and Mm*cpn60.1*-DPM.
Mtb*cpn60.1* contain a natural HindIII site 531 bps upstream to the stop codon. A HindIII site was inserted in Mm*cpn60.1* at the same position (i.e., 531 bps upstream to the stop codon) in the plasmid pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-8RBS (E) (this mutation was done by Doaa Ghulam in our lab). Both the plasmids pTrc-Mtb*cpn10*-Mtb*cpn60.1*_DPM and pTrc-Mm*cpn10*-Mm*cpn60.1*-CO-DPM-8RBS contained HindIII site after the *cpn60.1* gene A, B). The HindIII fragments were swapped, and the resultant plasmids are shown in C and D. Only the chaperonin genes regions of the plasmids are shown in A, B, C, and D.

The complementation assay showed that the chaperonin ability of MtbCpn60.1-DPM was lost by swapping of its 177 amino acids from C-terminal with those of MmCpn60.1-CO-DPM (Figure 7.16, Lane 7). MmCpn60.1_MtbHindIII_DPM could not perform the functional chaperonin activity (Figure 7.16, Lane 8). The loss of function by MtbCpn60.1_MmHindIII_DPM was surprising. We observed two results, one that putting the MtbCpn60.1 C-terminus into MmCpn60.1 doesn't enable it to complement. This means that the C terminus is not the important region, because putting in the C terminus from a functioning protein did not cause the MmCpn60.1-DPM to function as a chaperone. Second, putting the C-terminus from MmCpn60.1 into MtbCpn60.1 stopped it from functioning. This of course implies the reverse that the C-terminus is the important region. One possibility is that the C-terminus has a strong effect on protein expression but that can't be a complete explanation because we have seen considerable expression of MmCpn60.1-DPM-8RBS on western. Another possibility is that there is a specific interaction between a residue or region in the swapped fragment, and a residue or region elsewhere on the protein, and this is needed for function. In other words - without worrying about mechanism - you need "function" in both the swapped fragment and the unswapped fragment. MtbCpn60.1 has it both (that's why it works in *E. coli*); MmCpn60.1 has it in neither, that's why neither swap works

To find out that in what part of the swapped HindIII fragment does this important residue/region reside, a project student Gabriel Wong in our lab swapped only the last 13 residues of Mtbcpn60.1-DPM with the corresponding last 14 residues of MmCpn60.1-CO-DPM which also resulted in loss of chaperonin function of MtbCpn60.1-DPM. Gabriel’s results show that the last 13 residues are critical (as the results were same as seen in Figure 7.16). So, from this we know that something about that last 13 amino acids is important in explaining why MmCpn60.1 does not work. But it's not enough to explain why it doesn't work (because if it were, the addition of the MtbCpn60.1 residues would make it work in both experiments (Figure 7.16 and Gabriel's experiments).

This also demonstrates the importance of the diversity of C-terminal region of both the proteins. Out of the last 13 C-terminal residues, there are only 4 mismatches; MtbCpn60.1 contains lysine (K), alanine (A), histidine (H) and aspartic acid (D)

whereas at corresponding positions MmCpn60.1 contains glutamic acid (E), proline (P), aspartic acid (D) and glycine (G). MmCpn60.1 also contain one extra amino acid (glycine, G) in this region. MtbCpn60.1 C-terminal has more basic amino acids like lysine (K) and histidine (H) whereas at the corresponding sites MmCpn60.1 have acidic amino acids i.e., glutamic acid (E) and aspartic acid (D). The C-terminal of MmCpn60.1 seems to be more hydrophobic (more glycine, proline, and alanine residues) and more acidic.

Further investigation to find the least number of amino acids swapped from C-terminal which could cause loss of function of MtbCpn60.1-DPM would be advantageous. But it is not just about the C terminus, otherwise MmCpn60.1 with the Mtb terminus would work.

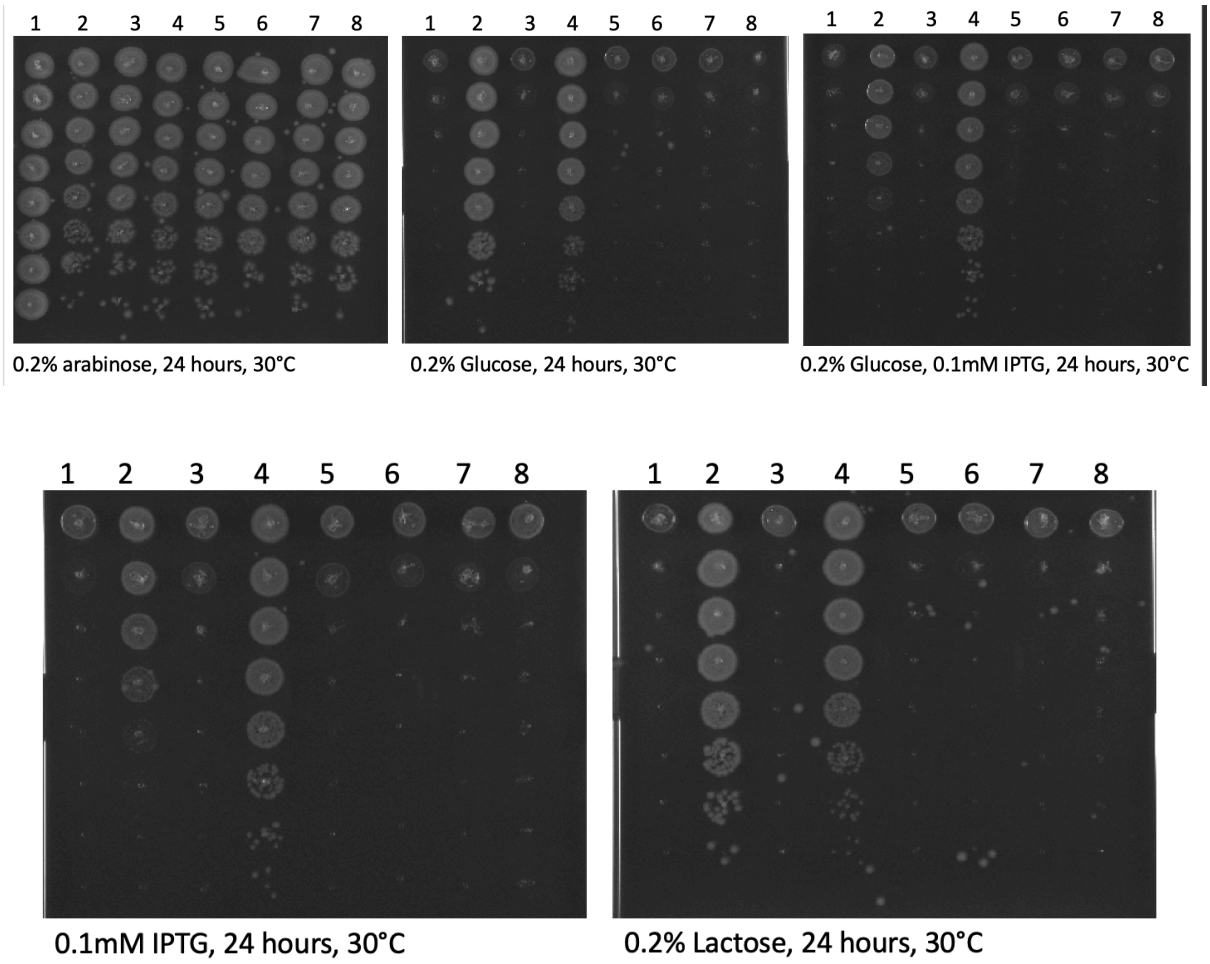


Figure 7.16 Complementation assay in MGM100.
Complementation assay in MGM100 harbouring plasmids 1 -8. All plates contained kanamycin and ampicillin. All images were taken after 24 hours of incubation at 30°C. A longer incubation up to 45 hours did not show any change in growth. 1) pTrc99A, 2) pTrc-*groESL*, 3) pTrc-Mtb*cpn10*-Mtb*cpn60.1* WT, 4) pTrc-Mtb*cpn10*-Mtb*cpn60.1*-DPM, 5) pTrc-Mm*cpn10*-Mm*cpn60.1*-CO (WT), 6) pTrc-Mmcpn10-Mmcpn60.1-CO-DPM-8RBS, 7) pTrc-Mtb*cpn10*-Mtb*cpn60.1*-MmHindIII-DPM, and 8) pTrc-Mm*cpn10*-Mm*cpn60.1*-MtbHindIII-DPM.

7.4 Mass spectrometry analysis

As discussed in the section 7.3.4 that the double phosphomimetic mutant of MmCpn60.1-CO did not show chaperone function but the same DPM of MtbCpn60.1 was able to complement the loss of GroEL in *E. coli*. Various other combinations of potential phosphate sites were also tested but MmCpn60.1 failed to complement the loss of GroEL in all our experiments. One reason could be that the MmCpn60.1 possess some species-specific phosphate sites. To find the post-translational modifications including phosphate sites of MmCpn60.1, we decided to purify the protein from *M. marinum* WT cultures and perform the mass spectrometry analysis.

7.4.1 Protein purification

MmCpn60.1 is hard to purify from the WT cultures because of its low expression levels and less stability. Two methods were used for the purification of MmCpn60.1. Taking the advantage of histidine rich C-terminal, Ni-NTA magnetic beads were used in first method. In second method, immunoprecipitation was performed using the antibody binding specifically to the C-terminal of MmCpn60.1. We confirmed that the antibody did not bind with MmCpn60.2 by wester blotting. The detailed procedure of Ni-NTA column purification and immunoprecipitation is given in chapter 2, section 2.7.4. 20ml of the purified proteins was loaded on two SDS-PAGE gels. One gel was stained whereas the second gel was used for western blotting. On the Coomassie Blue stained gel, the proteins were visible but not the level that could be captured in photograph. On western blot, a very clear band of MmCpn60.1 was seen (Western blot of MmCpn60.1 purified using immunoprecipitation and Ni-NTA magnetic beads.). In the lanes of IP_1 and IP_3 (two samples purified by immunoprecipitation) a very clear band of MmCpn60.1 was present, but a large band of cross-reactive proteins was also seen under than band of MmCpn60.1. In Ni-NTA the only one clear band of MmCpn60.1 was seen. Both the samples were sent to the Advanced Mass Spectrometry Facility at the School of Biosciences, University of Birmingham.

7.5 LC-MS/MS analysis

The protein samples were digested using trypsin, desalted and MS analysis was performed. The results of the first attempt of MS analysis did not show any post-translational modification positions in the MmCpn60.1 (Figure 7.18). Three peptides

having the S393 site (the corresponding site showed phosphorylation in MtbCpn60.1) are boxed in red. One reason for no phosphate sites observed in MS analysis could be that the phosphorylated protein might be present in very small amounts in the cell. Therefore, phosphate enrichment should be performed before subjecting the protein to the MS analysis for identification of post-translational modifications.

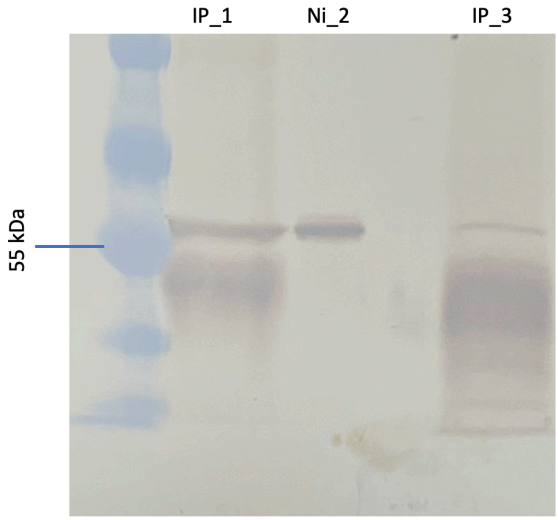


Figure 7.17 Western blot of MmCpn60.1 purified using immunoprecipitation and Ni-NTA magnetic beads.

Master Protein	Positions in Master Proteins	Confidence	Sequence in Protein	Positions in Protein	Peptide Group Modifications in Protein
MmCpn60.1	MmCpn60.1 [122-140]	High	K.AADAVSEALLASATPVSGK.D	[122-140]	
MmCpn60.1	MmCpn60.1 [379-389]	High	K.VGAATETELKK.R	[379-389]	
MmCpn60.1	MmCpn60.1 [238-266]	High	K.VAESGKPLMIIAEDIEGEALATLVVNSIR.K	[238-266]	
MmCpn60.1	MmCpn60.1 [238-266]	High	K.VAESGKPLMIIAEDIEGEALATLVVNSIR.K	[238-266]	
MmCpn60.1	MmCpn60.1 [80-100]	High	K.TNDVAGDGTTTATVLAQALVK.T	[80-100]	
MmCpn60.1	MmCpn60.1 [80-100]	High	K.TNDVAGDGTTTATVLAQALVK.T	[80-100]	
MmCpn60.1	MmCpn60.1 [42-57]	High	K.SFGGPTVTNDGVTVAR.D	[42-57]	
MmCpn60.1	MmCpn60.1 [501-511]	High	R.SAVLNASSVAR.M	[501-511]	
MmCpn60.1	MmCpn60.1 [501-511]	High	R.SAVLNASSVAR.M	[501-511]	
MmCpn60.1	MmCpn60.1 [320-343]	High	R.RVVVSKDDTIIVDGGGAPEAVEAR.V	[320-343]	
MmCpn60.1	MmCpn60.1 [390-402]	High	K.RKESVEDAVAAAK.A	[390-402]	
MmCpn60.1	MmCpn60.1 [512-539]	High	R.MVLTTETAIVDKPAEPEDDGHGHHGHAH.-	[512-539]	
MmCpn60.1	MmCpn60.1 [167-195]	High	K.VGHDGVVSVEESSTLGTELEFTEGVGFDK.G	[167-195]	
MmCpn60.1	MmCpn60.1 [512-539]	High	R.MVLTTETAIVDKPAEPEDDGHGHHGHAH.-	[512-539]	
MmCpn60.1	MmCpn60.1 [370-378]	High	K.LAGGVAVIK.V	[370-378]	
MmCpn60.1	MmCpn60.1 [391-402]	High	R.KESVEDAVAAAK.A	[391-402]	
MmCpn60.1	MmCpn60.1 [284-308]	High	R.KAFLQDLAAVTGAEVVNPDAGLVLR.E	[284-308]	
MmCpn60.1	MmCpn60.1 [196-224]	High	K.GYLSAYFVTDFDAQQAVLEDPLILLHQDK.I	[196-224]	
MmCpn60.1	MmCpn60.1 [153-166]	High	R.DQLIGDLVGEAMSK.V	[153-166]	
MmCpn60.1	MmCpn60.1 [326-343]	High	K.DDTIIVDGGGAPEAVEAR.V	[326-343]	
MmCpn60.1	MmCpn60.1 [429-467]	High	R.ASLSGDEAVGVDVFSEALAAPLYWIATNAGLDGSVVVNK.V	[429-467]	
MmCpn60.1	MmCpn60.1 [14-27]	High	R.AMEAGVDKLADTVR.V	[14-27]	
MmCpn60.1	MmCpn60.1 [285-308]	High	K.AFLQDLAAVTGAEVVNPDAGLVLR.E	[285-308]	
MmCpn60.1	MmCpn60.1 [403-421]	High	K.AAVEEGIVAGGGSALIQAR.N	[403-421]	
MmCpn60.1	MmCpn60.1 [122-152]	High	K.AADAVSEALLASATPVSGKDAIAQVATVSSR.D	[122-152]	
MmCpn60.1	MmCpn60.1 [105-121]	High	R.LVAAGINPIALGSGIGK.A	[105-121]	
MmCpn60.1	MmCpn60.1 [468-497]	High	K.VSELPAGHGLNAATLTYGDLAADGIVDPVK.V	[468-497]	

Figure 7.18 Results of Mass Spectrometry analysis of MmCpn60.1.

7.6 Discussion

Work done by Dr CM Santosh Kumar in our lab showed that a double phosphomimetic mutant of MtbCpn60.1-DPM (S75d+S393D) showed chaperone function in MGM100 when chromosomal copy of *groESL* was tightly supressed. This activity was lost when the ATPase activity of MtbCpn60.1-DPM was removed by SDM (ATPase mutations) as expected if the activity was indeed a result of Cpn60.1-DPM forming a functional chaperonin. One possible reason for complementation ability of MtbCpn60.1-DPM could be that the aspartic acid residues at 75 and 393 positions mimic the phosphoserine imitating a phosphorylated protein all the time which might facilitate the higher order oligomerization of the chaperonin.

Eleven unique "eukaryotic like" serine/threonine protein kinases (STPKs) of *M. tuberculosis* are known. The CDS sequences encoding the ser/thr protein kinases in *M. marinum* were searched and the sequence alignment showed a strong homology with most of the STPKs of *M. tuberculosis*. PknF of H37Rv has a corresponding CDS in 1218R (MM1218R_RS13070) with high sequence similarity and identity. Further experiments to investigate the role of STPKs in the phosphorylation of Cpn60.1 in mycobacteria could be used to test their role in the oligomerization story of mycobacterial chaperonin Cpn60.1. It was thought that because the STPKs of both the mycobacterial species have high sequence homology therefore the phosphorylation profiles of Cpn60.1 of both the species might be similar. To investigate this hypothesis, phosphomimetic mutations were constructed in Mm*cpn60.1*-CO at corresponding residues of MtbCpn60.1 that had been identified by Dr Santosh Kumar.

Various single, double, and triple phosphomimetic mutants of Mm*cpn60.1*-CO were constructed and tested for complementation, but none showed complementation phenotype in MGM100 in combination with the cognate co-chaperonin Mm*cpn10*. We therefore tested whether complementation was seen when the mutants were co-expressed with GroES. Unexpectedly, some complementation was seen in all cases, but also with the wild-type protein and when GroES was over-expressed on its own

The growth of all the strains harbouring pTrc derivative plasmids containing *groES* only and Mm*cpn60.1*-CO chaperonin gene in combination with *groES* irrespective of the phosphomimetic or ATPase mutation demonstrates that the maintenance of partial

growth might not be due to the chaperonin activity of MmCpn60.1. The growth of the strains harbouring plasmid having frameshift mutation in Mm*cpn60.1*-CO gene also showed similar results which again demonstrated that even in case the non-functional MmCpn60.1, the partial growth was sustained up to 10^{-4} dilutions. The growth could be due to the presence of excessive GroES getting expressed from plasmid but the exact reason for it is not known. I have not tested T25 and T54 positions of Canova et al., 2009, phosphomimetic mutations at these sites could also be tested in future experiments.

Multiple sequence alignment of Cpn60 homologues of ten mycobacterial species demonstrated that the C-termini of Cpn60.1 and Cpn60.2 were remarkably diverse. The C-terminal diversity was also observed among Cpn60.1 proteins from various mycobacterial species. Swapping of HindIII fragment could not cause MmCpn60.1-MtbHindIII-DPM to work as a functional chaperonin but it also led to loss of chaperonin ability of MtbCpn60.1-DPM. HindIII fragment consisted of 531 bps (encoding 177 amino acids). Swapping of last 13 residues of MtbCpn60.1-DPM with last 14 residues of MmCpn60.1 by Gabriel Wong (only a change of 4 amino acids with one addition of glycine) caused loss of chaperonin activity of MtbCpn60.1-DPM. This could be due to some specific interaction between a residue or region in the swapped fragment, and a residue or region elsewhere on the protein which was lost by swapping the last 13 residues.

Overall, the results of this part of the study demonstrated that the phosphorylation profiles of mycobacterial Cpn60.1 chaperonins might be species dependent. To look at the post-translational modifications of MmCpn60.1, mass spectrometry analysis of the purified MmCpn60.1 was performed. No post-translational modifications were observed in Mass spectrometry analysis in first attempt. In future attempts, phosphate enrichment of the purified protein might give better results.

8 Conclusion

The aim of this study was to understand the properties and functions of chaperonins of *M. marinum*. Mycobacterial chaperonins are interesting because of their immunogenic roles and many other moonlighting activities. Especially, Cpn60.1 was found to be essential for biofilm formation and for granulomatous inflammation of the infected mice.

The upstream regions of chaperonin genes of *M. marinum* were investigated for promoter activity. The strongest promoter activity was observed in the genomic fragment upstream to the Mm*cpn60.2*. Strong promoter activity was also found in the fragments upstream to the Mm*cpn10*, but no promoter activity was found in the intergenic region of Mm*cpn10*-Mm*cpn60.1*. The results are consistent with the previous findings that mycobacterial *cpn10-cpn60.*1 have bicistronic nature.

RNA-seq analysis of WT, ∆cpn60.1, and complemented strains of *M. marinum* was performed to study the transcriptomic profiles. The TPM values of read counts of the WT samples were used to observe the expression levels of chaperonin genes of *M. marinum* under normal conditions. Mm*cpn60.2* was found to be expressed more than double the levels of Mm*cpn10* and Mm*cpn10* expression levels were nearly doubled the expression levels of Mm*cpn60.1*. This indicates that despite of being in a single operon, the *cpn10* and *cpn60.1* are expressed differentially. The differential expression of Mm*cpn10* and Mm*cpn60.1* observed in our study is consistent with the previous findings.

We studied for the first time the effect of deletion of MmCpn60.1 on the transcriptomic profiles of the whole cell. The RNA-seq data was highly reproducible, quantitative, and precise. Many genes significantly upregulated in KO strain were reversed (downregulated) in the COM strains and vice versa. The gene ontology enrichment analysis of the common genes significantly upregulated in KO but downregulated in COM was performed. The top (according to fold enrichment) three important sets of pathways were found to be the key mechanisms of the cell.

The highest fold enrichment set of pathways was linked to the stress responses, chaperone activity, and protein folding. This shows that MmCpn60.1 might perform some chaperone function in its native cellular environment and that the deletion of *cpn60.1* caused an upregulation of *cpn60.*2 to compensate the protein folding activity.

A second reason could be that Cpn60.1 might be indirectly involved in some specific homeostasis activities (related to chaperone function and stress response) and that the absence of Cpn60.1 caused an increase in the expression levels of various chaperones to compensate the loss of Cpn60.1 activities.

The second highest fold enrichment set of pathways was linked to the cell wall, cell division, and peptidoglycan biogenesis. Several key genes (*mur* genes) involved in the peptidoglycan biosynthesis were upregulated in KO and downregulated in COM, with nearly same values of log2FoldChange. Cpn60.1 might play direct or indirect role in the cell wall and peptidoglycan maintenance and the absence of Cpn60.1 might had caused faster degradation of cell wall peptidoglycan, which in turn had stimulated the significant upregulation of the genes involved in the peptidoglycan biogenesis.

The third highest fold enrichment set of pathways observed was linked to the polyketide synthase pathway. The significant upregulation of two genes encoding the key enzymes, FadD22 and PKS 15/1, in KO strain involved in the biosynthesis of cell wall PGLs and DIMs is interesting. Both the enzymes catalyse the first two steps of PGL and DIM biosynthesis. Several enzymes involved in the transport of complex cell wall lipids across the cell membrane were also significantly upregulated in KO.

The observations of RNA-seq data suggest that although not essential for cell survival, MmCpn60.1 is involved in some key mechanisms of the cell which are of great significance in the infection. Cpn60.1 might be indirectly involved in the pathogenesis of tuberculosis by maintaining the cell wall integrity and performing cell homeostasis in the hostile environment of granulomas.

MmCpn60.1 could not support *E. coli* growth in the absence of GroEL both in combination with GroES and MmCpn10. The “sequence optimized” and RBS distance optimized MmCpn60.1 showed the similar results. The double phosphomimetic mutant (S75D & S393D) of MtbCpn60.1 complemented GroEL (experiments of Dr CM Santosh Kumar). However, the double phophomimetic mutant at the same sites in MmCpn60.1 could not complement GroEL, despite of detectable expression levels of MmCpn60.1-DPM-8RBS on SDS-PAGE gel. This could be due to several reason.

First, MmCpn60.1 might not be involved in protein folding which seems to be incorrect as we have observed in RNA-seq analysis data. Second, MmCpn60.1 might have species specific phosphorylation sites. Third, the C-terminal region is also important in some way because swapping of the C-terminal of MmCpn60.1 with that of MtbCpn60.1-DPM caused the loss of function of MtbCpn60.1-DPM. A fourth reason could be that the client proteins of MmCpn60.1 are different than GroEL and that MmCpn60.1 performs some chaperone activity in its natural cellular environment but does not assist in the assembly of some essential proteins which are the clients of GroEL.

We performed the mass spectrometric analysis of the purified MmCpn60.1 but in the first attempt no post-translation modification was found. One possible reason could be that very less percentage of the protein is phosphorylated in the cells and phosphate enrichment of the purified protein might give better results in mass spectrometric analysis.

This study demonstrates that MmCpn60.1 is dispensable but is likely playing important role in some key mechanisms in the proteostasis and membrane biogenesis of the cell. Further, it is expressed at a very low levels, despite being in the same operon with Mm*cpn10*, Mm*cpn60.1* is expressed at levels two-fold lower than Mm*cpn10*, possibly because of the polar effects or post-transcriptional effects. Comprehensive investigation of the role of Cpn60.1 in the maintenance of mycobacterial cell wall would help the understanding of its role in the mycobacterial survival in the hostile environment of granulomas.

References

Agarwal, N., & Tyagi, A. K. (2006). Mycobacterial transcriptional signals: requirements for recognition by RNA polymerase and optimal transcriptional activity. *Nucleic Acids Research*, *34*(15), 4245–4257. https://doi.org/10.1093/nar/gkl521

Ahmed, M. T. (2010). *Functional and structural characterisation of Mycobacterial Chaperonins.* . University of Birmingham.

Amann, E., Ochs, B., & Abel, K.-J. (1988). Tightly regulated tac promoter vectors useful for the expression of unfused and fused proteins in Escherichia coli. *Gene*, *69*(2), 301–315. https://doi.org/10.1016/0378-1119(88)90440-4

Amsterdam, A., Burgess, S., Golling, G., Chen, W., Sun, Z., Townsend, K., Farrington, S., Haldi, M., & Hopkins, N. (1999). A large-scale insertional mutagenesis screen in zebrafish. *Genes & Development*, *13*(20), 2713–2724. https://doi.org/10.1101/gad.13.20.2713

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Andersen, P., & Doherty, T. M. (2005). The success and failure of BCG — implications for a novel tuberculosis vaccine. *Nature Reviews Microbiology*, *3*(8), 656–662. https://doi.org/10.1038/nrmicro1211

Anfinsen, C. B., Haber, E., Sela, M., & White, F. H. (1961). THE KINETICS OF FORMATION OF NATIVE RIBONUCLEASE DURING OXIDATION OF THE REDUCED POLYPEPTIDE CHAIN. *Proceedings of the National Academy of Sciences*, *47*(9), 1309–1314. https://doi.org/10.1073/pnas.47.9.1309

Ang, D., & Georgopoulos, C. (1989). The heat-shock-regulated grpE gene of Escherichia coli is required for bacterial growth at all temperatures but is dispensable in certain mutant backgrounds. *Journal of Bacteriology*, *171*(5), 2748–2755. https://doi.org/10.1128/jb.171.5.2748-2755.1989

Aravindhan, V., Christy, A. J., Roy, S., Ajitkumar, P., Narayanan, P. R., & Narayanan, S. (2009). Mycobacterium tuberculosis groE promoter controls the expression of the bicistronic groESL1 operon and shows differential regulation under stress conditions. *FEMS Microbiology Letters*, *292*(1), 42–49. https://doi.org/10.1111/j.1574-6968.2008.01465.x

Astin, J. W., Keerthisinghe, P., Du, L., Sanderson, L. E., Crosier, K. E., Crosier, P. S., & Hall, C. J. (2017). *Innate immune cells and bacterial infection in zebrafish* (pp. 31–60). https://doi.org/10.1016/bs.mcb.2016.08.002

Av-Gay, Y., & Everett, M. (2000). The eukaryotic-like Ser/Thr protein kinases of Mycobacterium tuberculosis. *Trends in Microbiology*, *8*(5), 238–244. https://doi.org/10.1016/S0966-842X(00)01734-0

Azoulay, I., Kucherenko, N., Nachliel, E., Gutman, M., Azem, A., & Tsfadia, Y. (2013). Tracking the Interplay between Bound Peptide and the Lid Domain of DnaK, Using Molecular Dynamics. *International Journal of Molecular Sciences*, *14*(6), 12675–12695. https://doi.org/10.3390/ijms140612675

Bachmann, N. L., Salamzade, R., Manson, A. L., Whittington, R., Sintchenko, V., Earl, A. M., & Marais, B. J. (2020). Key Transitions in the Evolution of Rapid and Slow

Growing Mycobacteria Identified by Comparative Genomics. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.03019

Balabanova, Y., Ignatyeva, O., Fiebig, L., Riekstina, V., Danilovits, M., Jaama, K., Davidaviciene, E., Radiulyte, B., Popa, C. M., Nikolayevskyy, V., & Drobniewski, F. (2016). Survival of patients with multidrug-resistant TB in Eastern Europe: what makes a difference? *Thorax*, *71*(9), 854–861. https://doi.org/10.1136/thoraxjnl-2015-207638

Balchin, D., Hayer-Hartl, M., & Hartl, F. U. (2020). Recent advances in understanding catalysis of protein folding by molecular chaperones. *FEBS Letters*, *594*(17), 2770–2781. https://doi.org/10.1002/1873-3468.13844

Barberis, I., Bragazzi, N. L., Galluzzo, L., & Martini, M. (2017). The history of tuberculosis: from the first historical records to the isolation of Koch's bacillus. In *J PREV MED HYG* (Vol. 58).

Barends, T. R., Werbeck, N. D., & Reinstein, J. (2010). Disaggregases in 4 dimensions. *Current Opinion in Structural Biology*, *20*(1), 46–53. https://doi.org/10.1016/j.sbi.2009.12.014

Barreiro, C., González-Lavado, E., Brand, S., Tauch, A., & Martín, J. F. (2005). Heat shock proteome analysis of wild-type Corynebacterium glutamicum ATCC 13032 and a spontaneous mutant lacking GroEL1, a dispensable chaperone. *Journal of Bacteriology*, *187*(3), 884–889. https://doi.org/10.1128/JB.187.3.884-889.2005

Barreiro, C., González-Lavado, E., Pátek, M., & Martín, J. F. (2004). Transcriptional analysis of the groES-groEL1, groEL2, and dnaK genes in Corynebacterium glutamicum: Characterization of heat shock-induced promoters. *Journal of Bacteriology*, *186*(14), 4813–4817. https://doi.org/10.1128/JB.186.14.4813-4817.2004

Bartolucci, C., Lamba, D., Grazulis, S., Manakova, E., & Heumann, H. (2005). Crystal Structure of Wild-type Chaperonin GroEL. *Journal of Molecular Biology*, *354*(4), 940–951. https://doi.org/10.1016/j.jmb.2005.09.096

Batt, S. M., Minnikin, D. E., & Besra, G. S. (2020). The thick waxy coat of mycobacteria, a protective layer against antibiotics and the host's immune system. *Biochemical Journal*, *477*(10), 1983–2006. https://doi.org/10.1042/BCJ20200194

Becker, K., & Sander, P. (2016). *Mycobacterium tuberculosis* lipoproteins in virulence and immunity - fighting with a double-edged sword. *FEBS Letters*, *590*(21), 3800–3819. https://doi.org/10.1002/1873-3468.12273

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berg, S., Kaur, D., Jackson, M., & Brennan, P. J. (2007). The glycosyltransferases of Mycobacterium tuberculosis—roles in the synthesis of arabinogalactan, lipoarabinomannan, and other glycoconjugates. *Glycobiology*, *17*(6), 35R-56R. https://doi.org/10.1093/glycob/cwm010

Berglund, E. C., Kiialainen, A., & Syvänen, A.-C. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative Genetics*, *2*(1), 23. https://doi.org/10.1186/2041-2223-2-23

Bergonzelli, G. E., Granato, D., Pridmore, R. D., Marvin-Guy, L. F., Donnicola, D., & Corthésy-Theulaz, I. E. (2006). GroEL of *Lactobacillus johnsonii* La1 (NCC 533) Is Cell Surface Associated: Potential Role in Interactions with the Host and the

Gastric Pathogen *Helicobacter pylori*. *Infection and Immunity*, *74*(1), 425–434. https://doi.org/10.1128/IAI.74.1.425-434.2006
Bhargavi, G., Singh, A. K., Patil, S. A., & Palaniyandi, K. (2022). A putative short-chain dehydrogenase Rv0148 of Mycobacterium tuberculosis affects bacterial survival and virulence. *Current Research in Microbial Sciences*, *3*, 100113. https://doi.org/10.1016/j.crmicr.2022.100113
Blackburn, P. R., Campbell, J. M., Clark, K. J., & Ekker, S. C. (2013). The CRISPR System—Keeping Zebrafish Gene Targeting Fresh. *Zebrafish*, *10*(1), 116–118. https://doi.org/10.1089/zeb.2013.9999
Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170
Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519
Brocchieri, L., & Karlin, S. (2008). Conservation among HSP60 sequences in relation to structure, function, and evolution. *Protein Science*, *9*(3), 476–486. https://doi.org/10.1110/ps.9.3.476
Bross, P., Magnoni, R., & Sigaard Bie, A. (2013). Molecular Chaperone Disorders: Defective Hsp60 in Neurodegeneration. *Current Topics in Medicinal Chemistry*, *12*(22), 2491–2503. https://doi.org/10.2174/1568026611212220005
Bukau, B., & Horwich, A. L. (1998). The Hsp70 and Hsp60 Chaperone Machines. *Cell*, *92*(3), 351–366. https://doi.org/10.1016/S0092-8674(00)80928-9
Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, *11*(1), 94. https://doi.org/10.1186/1471-2105-11-94
Campbell, J. M., Hartjes, K. A., Nelson, T. J., Xu, X., & Ekker, S. C. (2013). New and TALENted Genome Engineering Toolbox. *Circulation Research*, *113*(5), 571–587. https://doi.org/10.1161/CIRCRESAHA.113.301765
Canova, M. J., Kremer, L., & Molle, V. (2009). The Mycobacterium tuberculosis GroEL1 Chaperone Is a Substrate of Ser/Thr Protein Kinases. *Journal of Bacteriology*, *191*(8), 2876–2883. https://doi.org/10.1128/JB.01569-08
Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M.-A., Barrell, B. G., & Parkhill, J. (2005). ACT: the Artemis comparison tool. *Bioinformatics*, *21*(16), 3422–3423. https://doi.org/10.1093/bioinformatics/bti553
Cassio Barreto de Oliveira, M., & Balan, A. (2020). The ATP-Binding Cassette (ABC) Transport Systems in Mycobacterium tuberculosis: Structure, Function, and Possible Targets for Therapeutics. *Biology*, *9*(12), 443. https://doi.org/10.3390/biology9120443
Castanié-Cornet, M.-P., Bruel, N., & Genevaux, P. (2014). Chaperone networking facilitates protein targeting to the bacterial cytoplasmic membrane. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1843*(8), 1442–1456. https://doi.org/10.1016/j.bbamcr.2013.11.007
Chakraborty, P., Bajeli, S., Kaushal, D., Radotra, B. D., & Kumar, A. (2021). Biofilm formation in the lung contributes to virulence and drug tolerance of Mycobacterium tuberculosis. *Nature Communications*, *12*(1), 1606. https://doi.org/10.1038/s41467-021-21748-6

Chang, J. C., Jagirdar, J., & Lesser, M. (1986). Long-term evolution of BCG- and CFA-induced granulomas in rat lungs. Correlation of histologic features with cells in bronchoalveolar lavage samples. *The American Journal of Pathology*, *125*(1), 16–27.

Chełkowska-Pauszek, A., Kosiński, J. G., Marciniak, K., Wysocka, M., Bakowska-żywicka¸, K., & Żywicki, M. (2021). The role of rna secondary structure in regulation of gene expression in bacteria. In *International Journal of Molecular Sciences* (Vol. 22, Issue 15). MDPI AG. https://doi.org/10.3390/ijms22157845

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., Zhang, X., Wang, J., Yang, H., Fang, L., & Chen, Q. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience*, *7*(1). https://doi.org/10.1093/gigascience/gix120

Chiaruttini, C., & Guillier, M. (2019). *On the role of mRNA secondary structure in bacterial translation Article Title: On the role of mRNA secondary structure in bacterial translation No conflict of interest*. https://doi.org/10.1002/wrna.1579ï

Chung-Delgado, K., Guillen-Bravo, S., Revilla-Montag, A., & Bernabe-Ortiz, A. (2015). Mortality among MDR-TB Cases: Comparison with Drug-Susceptible Tuberculosis and Associated Factors. *PLOS ONE*, *10*(3), e0119332. https://doi.org/10.1371/journal.pone.0119332

Clark, H. F., & Shepard, C. C. (1963). EFFECT OF ENVIRONMENTAL TEMPERATURES ON INFECTION WITH *MYCOBACTERIUM MARINUM* (BALNEI) OF MICE AND A NUMBER OF POIKILOTHERMIC SPECIES. *Journal of Bacteriology*, *86*(5), 1057–1069. https://doi.org/10.1128/jb.86.5.1057-1069.1963

Coates, A. R. M., Shinnick, T. M., & Ellis, R. J. (1993). Chaperonin nomenclature. *Molecular Microbiology*, *8*(4), 787–787. https://doi.org/10.1111/j.1365-2958.1993.tb01624.x

Cohen, P. (2002). The origins of protein phosphorylation. *Nature Cell Biology*, *4*(5), E127–E130. https://doi.org/10.1038/ncb0502-e127

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Cook, G. M., Berney, M., Gebhard, S., Heinemann, M., Cox, R. A., Danilchanka, O., & Niederweis, M. (2009). *Physiology of Mycobacteria* (pp. 81–319). https://doi.org/10.1016/S0065-2911(09)05502-7

Corrales, F. J., & Fersht, A. R. (1996). Toward a mechanism for GroEL.GroES chaperone activity: an ATPase-gated and -pulsed folding and annealing cage. *Proceedings of the National Academy of Sciences*, *93*(9), 4509–4512. https://doi.org/10.1073/pnas.93.9.4509

Croucher, N. J., & Thomson, N. R. (2010). Studying bacterial transcriptomes using RNA-seq. *Current Opinion in Microbiology*, *13*(5), 619–624. https://doi.org/10.1016/j.mib.2010.09.009

Daegelen, P., Studier, F. W., Lenski, R. E., Cure, S., & Kim, J. F. (2009). Tracing Ancestors and Relatives of Escherichia coli B, and the Derivation of B Strains REL606 and

BL21(DE3). *Journal of Molecular Biology*, *394*(4), 634–643. https://doi.org/10.1016/j.jmb.2009.09.022
d'Andrea, F. B., Poulton, N. C., Froom, R., Tam, K., Campbell, E. A., & Rock, J. M. (2022). The essential *M. tuberculosis* Clp protease is functionally asymmetric in vivo. *Science Advances*, *8*(18). https://doi.org/10.1126/sciadv.abn7943
Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008
Daniel, T. M. (2006). The history of tuberculosis. *Respiratory Medicine*, *100*(11), 1862–1870. https://doi.org/10.1016/j.rmed.2006.08.006
Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research*, *14*(7), 1394–1403. https://doi.org/10.1101/gr.2289704
Das, S., Pettersson, B. M. F., Behra, P. R. K., Mallick, A., Cheramie, M., Ramesh, M., Shirreff, L., DuCote, T., Dasgupta, S., Ennis, D. G., & Kirsebom, Leif. A. (2018). Extensive genomic diversity among Mycobacterium marinum strains revealed by whole genome sequencing. *Scientific Reports*, *8*(1), 12040. https://doi.org/10.1038/s41598-018-30152-y
Davis, J. M., & Ramakrishnan, L. (2009). The Role of the Granuloma in Expansion and Dissemination of Early Tuberculous Infection. *Cell*, *136*(1), 37–49. https://doi.org/10.1016/j.cell.2008.11.014
de Bruyn, J., Bosmans, R., Turneer, M., Weckx, M., Nyabenda, J., van Vooren, J. P., Falmagne, P., Wiker, H. G., & Harboe, M. (1987). Purification, partial characterization, and identification of a skin-reactive protein antigen of Mycobacterium bovis BCG. *Infection and Immunity*, *55*(1), 245–252. https://doi.org/10.1128/iai.55.1.245-252.1987
de León, P., Marco, S., Isiegas, C., Marina, A., Carrascosa, J. L., & Mellado, R. P. (1997). Streptomyces lividans groES, groEL1 and groEL2 genes. *Microbiology*, *143*(11), 3563–3571. https://doi.org/10.1099/00221287-143-11-3563
Delogu, G., Sali, M., & Fadda, G. (2013). THE BIOLOGY OF MYCOBACTERIUM TUBERCULOSIS INFECTION. *Mediterranean Journal of Hematology and Infectious Diseases*, *5*(1), e2013070. https://doi.org/10.4084/mjhid.2013.070
DeLuca, D. S., Levin, J. Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W., & Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, *28*(11), 1530–1532. https://doi.org/10.1093/bioinformatics/bts196
Deol, P., Vohra, R., Saini, A. K., Singh, A., Chandra, H., Chopra, P., Das, T. K., Tyagi, A. K., & Singh, Y. (2005). Role of *Mycobacterium tuberculosis* Ser/Thr Kinase PknF: Implications in Glucose Transport and Cell Division. *Journal of Bacteriology*, *187*(10), 3415–3420. https://doi.org/10.1128/JB.187.10.3415-3420.2005
Deuerling, E., Schulze-Specking, A., Tomoyasu, T., Mogk, A., & Bukau, B. (1999). Trigger factor and DnaK cooperate in folding of newly synthesized proteins. *Nature*, *400*(6745), 693–696. https://doi.org/10.1038/23301
Devulder, G., de Montclos, M. P., & Flandrois, J. P. (2005). A multigene approach to phylogenetic analysis using the genus Mycobacterium as a model. *International Journal of Systematic and Evolutionary Microbiology*, *55*(1), 293–302. https://doi.org/10.1099/ijs.0.63222-0

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dooley, S. W. (1992). Multidrug-resistant Tuberculosis. *Annals of Internal Medicine*, *117*(3), 257. https://doi.org/10.7326/0003-4819-117-3-257

Dreyfus, M., & Régnier, P. (2002). The Poly(A) Tail of mRNAs. *Cell*, *111*(5), 611–613. https://doi.org/10.1016/S0092-8674(02)01137-6

Driever, W., Solnica-Krezel, L., Schier, A. F., Neuhauss, S. C., Malicki, J., Stemple, D. L., Stainier, D. Y., Zwartkruis, F., Abdelilah, S., Rangini, Z., Belak, J., & Boggs, C. (1996). A genetic screen for mutations affecting embryogenesis in zebrafish. *Development*, *123*(1), 37–46. https://doi.org/10.1242/dev.123.1.37

Dyer, N. P., Shahrezaei, V., & Hebenstreit, D. (2019). LiBiNorm: an htseq-count analogue with improved normalisation of Smart-seq2 data and library preparation diagnostics. *PeerJ*, *7*, e6222. https://doi.org/10.7717/peerj.6222

Ehlers, S., & Schaible, U. E. (2013). The Granuloma in Tuberculosis: Dynamics of a Host–Pathogen Collusion. *Frontiers in Immunology*, *3*. https://doi.org/10.3389/fimmu.2012.00411

Ellis, J. (1987). Proteins as molecular chaperones. *Nature*, *328*(6129), 378–379. https://doi.org/10.1038/328378a0

Ellis, R. J. (1993). The general concept of molecular chaperones. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *339*(1289), 257–261. https://doi.org/10.1098/rstb.1993.0023

Ensgraber, M., & Loos, M. (1992). A 66-kilodalton heat shock protein of Salmonella typhimurium is responsible for binding of the bacterium to intestinal mucus. *Infection and Immunity*, *60*(8), 3072–3078. https://doi.org/10.1128/iai.60.8.3072-3078.1992

Fan, M., Rao, T., Zacco, E., Ahmed, M. T., Shukla, A., Ojha, A., Freeke, J., Robinson, C. v., Benesch, J. L., & Lund, P. A. (2012). The unusual mycobacterial chaperonins: evidence for *in vivo* oligomerization and specialization of function. *Molecular Microbiology*, *85*(5), 934–944. https://doi.org/10.1111/j.1365-2958.2012.08150.x

Fischer, H. M., Babst, M., Kaspar, T., Acuial, G., Arigoni2, F., & Hennecke, H. (1993). One member of a groESL-like chaperonin multigene family in Bradyrhizobium japonicum is co-regulated with symbiotic nitrogen fixation genes. In *The EMBO Journal* (Vol. 12, Issue 7).

Fischer, K., Chatterjee, D., Torrelles, J., Brennan, P. J., Kaufmann, S. H. E., & Schaible, U. E. (2001). Mycobacterial Lysocardiolipin Is Exported from Phagosomes upon Cleavage of Cardiolipin by a Macrophage-Derived Lysosomal Phospholipase A $_2$. *The Journal of Immunology*, *167*(4), 2187–2192. https://doi.org/10.4049/jimmunol.167.4.2187

Flynn, J. L., Chan, J., & Lin, P. L. (2011). Macrophages and control of granulomatous inflammation in tuberculosis. *Mucosal Immunology*, *4*(3), 271–278. https://doi.org/10.1038/mi.2011.14

FRIEDLAND, J. S., SHATTOCK, R., REMICK, D. G., & GRIFFIN, G. E. (2008). Mycobacterial 65-kD heat shock protein induces release of proinflammatory cytokines from human monocytic cells. *Clinical and Experimental Immunology*, *91*(1), 58–62. https://doi.org/10.1111/j.1365-2249.1993.tb03354.x

Garduño, R. A., Chong, A., Nasrallah, G. K., & Allan, D. S. (2011). The Legionella pneumophila Chaperonin – An Unusual Multifunctional Protein in Unusual

Locations. *Frontiers in Microbiology*, *2*. https://doi.org/10.3389/fmicb.2011.00122

Garnier, T., Eiglmeier, K., Camus, J.-C., Medina, N., Mansoor, H., Pryor, M., Duthoy, S., Grondin, S., Lacroix, C., Monsempe, C., Simon, S., Harris, B., Atkin, R., Doggett, J., Mayes, R., Keating, L., Wheeler, P. R., Parkhill, J., Barrell, B. G., … Hewinson, R. G. (2003). The complete genome sequence of *Mycobacterium bovis*. *Proceedings of the National Academy of Sciences*, *100*(13), 7877–7882. https://doi.org/10.1073/pnas.1130426100

Gentz, R., Langner, A., Chang, A. C., Cohen, S. N., & Bujard, H. (1981). Cloning and analysis of strong promoters is made possible by the downstream placement of a RNA termination signal. *Proceedings of the National Academy of Sciences of the United States of America*, *78*(8), 4936–4940. https://doi.org/10.1073/pnas.78.8.4936

GEORGOPOULOS, C. P., HENDRIX, R. W., KAISER, A. D., & WOOD, W. B. (1972). Role of the Host Cell in Bacteriophage Morphogenesis: Effects of a Bacterial Mutation on T4 Head Assembly. *Nature New Biology*, *239*(89), 38–41. https://doi.org/10.1038/newbio239038a0

Gey van Pittius, N. C., Sampson, S. L., Lee, H., Kim, Y., van Helden, P. D., & Warren, R. M. (2006a). Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. *BMC Evolutionary Biology*, *6*(1), 95. https://doi.org/10.1186/1471-2148-6-95

Gey van Pittius, N. C., Sampson, S. L., Lee, H., Kim, Y., van Helden, P. D., & Warren, R. M. (2006b). Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. *BMC Evolutionary Biology*, *6*(1), 95. https://doi.org/10.1186/1471-2148-6-95

Ghosh, S., & Chan, C.-K. K. (2016). *Analysis of RNA-Seq Data Using TopHat and Cufflinks* (pp. 339–361). https://doi.org/10.1007/978-1-4939-3167-5_18

Goloubinoff, P., Christeller, J. T., Gatenby, A. A., & Lorimer, G. H. (1989). Reconstitution of active dimeric ribulose bisphosphate carboxylase from an unfolded state depends on two chaperonin proteins and Mg-ATP. *Nature*, *342*(6252), 884–889. https://doi.org/10.1038/342884a0

Goloubinoff, P., Sassi, A. S., Fauvet, B., Barducci, A., & de Los Rios, P. (2018). Chaperones convert the energy from ATP into the nonequilibrium stabilization of native proteins. *Nature Chemical Biology*, *14*(4), 388–395. https://doi.org/10.1038/s41589-018-0013-8

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Goyal, K., Qamra, R., & Mande, S. C. (2006). Multiple Gene Duplication and Rapid Evolution in the groEL Gene: Functional Implications. *Journal of Molecular Evolution*, *63*(6), 781–787. https://doi.org/10.1007/s00239-006-0037-7

Gragerov, A., Nudler, E., Komissarova, N., Gaitanaris, G. A., Gottesman, M. E., & Nikiforov, V. (1992). Cooperation of GroEL/GroES and DnaK/DnaJ heat shock proteins in preventing protein misfolding in Escherichia coli. *Proceedings of the National Academy of Sciences*, *89*(21), 10341–10344. https://doi.org/10.1073/pnas.89.21.10341

Grandvalet, C., de Crecy-Lagard, V., & Mazodier, P. (1999). The ClpB ATPase of Streptomyces albus G belongs to the HspR heat shock regulon. *Molecular Microbiology*, *31*(2), 521–532. https://doi.org/10.1046/j.1365-2958.1999.01193.x

Grandvalet, C., Rapoport, G., & Mazodier, P. (1998). *hrcA* , Encoding the Repressor of the *groEL* Genes in *Streptomyces albus G* , Is Associated with a Second *dnaJ* Gene. *Journal of Bacteriology*, *180*(19), 5129–5134. https://doi.org/10.1128/JB.180.19.5129-5134.1998

Grandvalet, C., Servant, P., & Mazodier, P. (1997). *Disruption of hspR, the repressor gene of the dnaK operon in Streptomyces albus G*.

Grange, J. M., & Zumla, A. (2002). The global emergency of tuberculosis: what is the cause? *Journal of the Royal Society for the Promotion of Health*, *122*(2), 78–81. https://doi.org/10.1177/146642400212200206

Grant, G. R., Farkas, M. H., Pizarro, A. D., Lahens, N. F., Schug, J., Brunk, B. P., Stoeckert, C. J., Hogenesch, J. B., & Pierce, E. A. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, *27*(18), 2518–2528. https://doi.org/10.1093/bioinformatics/btr427

Grant, S. G., Jessee, J., Bloom, F. R., & Hanahan, D. (1990). Differential plasmid rescue from transgenic mouse DNAs into Escherichia coli methylation-restriction mutants. *Proceedings of the National Academy of Sciences*, *87*(12), 4645–4649. https://doi.org/10.1073/pnas.87.12.4645

Graubert, A., Aguet, F., Ravi, A., Ardlie, K. G., & Getz, G. (2021). RNA-SeQC 2: efficient RNA-seq quality control and quantification for large cohorts. *Bioinformatics*, *37*(18), 3048–3050. https://doi.org/10.1093/bioinformatics/btab135

Grossman, A. D., Straus, D. B., Walter, W. A., & Gross, C. A. (1987). *cr 32 synthesis can regulate the synthesis of heat shock proteins in Escherichia coli*.

Gupta, R. S., Lo, B., & Son, J. (2018). Phylogenomics and Comparative Genomic Studies Robustly Support Division of the Genus Mycobacterium into an Emended Genus Mycobacterium and Four Novel Genera. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.00067

Gutman, G. A., & Hatfield, G. W. (1989). Nonrandom utilization of codon pairs in Escherichia coli. *Proceedings of the National Academy of Sciences*, *86*(10), 3699–3703. https://doi.org/10.1073/pnas.86.10.3699

Guzman, L. M., Belin, D., Carson, M. J., & Beckwith, J. (1995). Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *Journal of Bacteriology*, *177*(14), 4121–4130. https://doi.org/10.1128/jb.177.14.4121-4130.1995

Hagemann, L., Gründel, A., Jacobs, E., & Dumke, R. (2017). The surface-displayed chaperones GroEL and DnaK of Mycoplasma pneumoniae interact with human plasminogen and components of the extracellular matrix. *Pathogens and Disease*, *75*(3). https://doi.org/10.1093/femspd/ftx017

Hartl, F. U., & Hayer-Hartl, M. (2002). Molecular Chaperones in the Cytosol: from Nascent Chain to Folded Protein. *Science*, *295*(5561), 1852–1858. https://doi.org/10.1126/science.1068408

Hawley, D. K., & McClure, W. R. (1983). Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucleic Acids Research*, *11*(8), 2237–2255. https://doi.org/10.1093/nar/11.8.2237

Hayer-Hartl, M., Bracher, A., & Hartl, F. U. (2016). The GroEL–GroES Chaperonin Machine: A Nano-Cage for Protein Folding. *Trends in Biochemical Sciences*, *41*(1), 62–76. https://doi.org/10.1016/j.tibs.2015.07.009

Hayer-Hartl, M. K., Martin, J., & Hartl, F. U. (1995). Asymmetrical Interaction of GroEL and GroES in the ATPase Cycle of Assisted Protein Folding. *Science*, *269*(5225), 836–841. https://doi.org/10.1126/science.7638601

Hecker, M., Schumann, W., & Vö Lker, U. (1996). *MicroReview Heat-shock and general stress response in Bacillus subtilis*.

Hemmingsen, S. M., Woolford, C., van der Vies, S. M., Tilly, K., Dennis, D. T., Georgopoulos, C. P., Hendrix, R. W., & Ellis, R. J. (1988). Homologous plant and bacterial proteins chaperone oligomeric protein assembly. *Nature*, *333*(6171), 330–334. https://doi.org/10.1038/333330a0

Henderson, B., Allan, E., & Coates, A. R. M. (2006). Stress Wars: the Direct Role of Host and Bacterial Molecular Chaperones in Bacterial Infection. *Infection and Immunity*, *74*(7), 3693–3706. https://doi.org/10.1128/IAI.01882-05

Henderson, B., Lund, P. A., & Coates, A. R. M. (2010). Multiple moonlighting functions of mycobacterial molecular chaperones. *Tuberculosis*, *90*(2), 119–124. https://doi.org/10.1016/j.tube.2010.01.004

Hitomi, M., Nishimura, H., Tsujimoto, Y., Matsui, H., & Watanabe, K. (2003). Identification of a helix-turn-helix motif of Bacillus thermoglucosidasius HrcA essential for binding to the CIRCE element and thermostability of the HrcA-CIRCE complex, indicating a role as a thermosensor. *Journal of Bacteriology*, *185*(1), 381–385. https://doi.org/10.1128/JB.185.1.381-385.2003

Hoffmann, A., Bukau, B., & Kramer, G. (2010). Structure and function of the molecular chaperone Trigger Factor. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1803*(6), 650–661. https://doi.org/10.1016/j.bbamcr.2010.01.017

Hoffmann, A., Merz, F., Rutkowska, A., Zachmann-Brand, B., Deuerling, E., & Bukau, B. (2006). Trigger Factor Forms a Protective Shield for Nascent Polypeptides at the Ribosome. *Journal of Biological Chemistry*, *281*(10), 6539–6545. https://doi.org/10.1074/jbc.M512345200

Hong, M., Tao, S., Zhang, L., Diao, L.-T., Huang, X., Huang, S., Xie, S.-J., Xiao, Z.-D., & Zhang, H. (2020a). RNA sequencing: new technologies and applications in cancer research. *Journal of Hematology & Oncology*, *13*(1), 166. https://doi.org/10.1186/s13045-020-01005-x

Hong, M., Tao, S., Zhang, L., Diao, L.-T., Huang, X., Huang, S., Xie, S.-J., Xiao, Z.-D., & Zhang, H. (2020b). RNA sequencing: new technologies and applications in cancer research. *Journal of Hematology & Oncology*, *13*(1), 166. https://doi.org/10.1186/s13045-020-01005-x

Horovitz, A., Bochkareva, E. S., Kovalenko, O., & Girshovich, A. S. (1993). Mutation Ala2 → Ser Destabilizes Intersubunit Interactions in the Molecular Chaperone GroEL. *Journal of Molecular Biology*, *231*(1), 58–64. https://doi.org/10.1006/jmbi.1993.1256

Horwich, A. L., Apetri, A. C., & Fenton, W. A. (2009). The GroEL/GroES *cis* cavity as a passive anti-aggregation device. *FEBS Letters*, *583*(16), 2654–2662. https://doi.org/10.1016/j.febslet.2009.06.049

Horwich, A. L., & Fenton, W. A. (2009). Chaperonin-mediated protein folding: using a central cavity to kinetically assist polypeptide chain folding. *Quarterly Reviews of Biophysics*, *42*(2), 83–116. https://doi.org/10.1017/S0033583509004764

Horwich, A. L., & Fenton, W. A. (2020). Chaperonin-assisted protein folding: a chronologue. *Quarterly Reviews of Biophysics*, *53*, e4. https://doi.org/10.1017/S0033583519000143

Horwitz, J. P., Chua, J., Curby, R. J., Tomson, A. J., da Rooge, M. A., Fisher, B. E., Mauricio, J., & Klundt, I. (1964). Substrates for Cytochemical Demonstration of Enzyme Activity. I. Some Substituted 3-Indolyl-β-D-glycopyranosides [1a]. *Journal of Medicinal Chemistry*, *7*(4), 574–575. https://doi.org/10.1021/jm00334a044

Houry, W. A., Frishman, D., Eckerskorn, C., Lottspeich, F., & Hartl, F. U. (1999). Identification of in vivo substrates of the chaperonin GroEL. *Nature*, *402*(6758), 147–154. https://doi.org/10.1038/45977

Howard, D., & Benson, K. (2003). Evolutionary computation method for pattern recognition of cis-acting sites. *Biosystems*, *72*(1–2), 19–27. https://doi.org/10.1016/S0303-2647(03)00132-1

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J. C., Koch, R., Rauch, G.-J., White, S., … Stemple, D. L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*(7446), 498–503. https://doi.org/10.1038/nature12111

Hu, Y., Henderson, B., Lund, P. A., Tormay, P., Ahmed, M. T., Gurcha, S. S., Besra, G. S., & Coates, A. R. M. (2008). A Mycobacterium tuberculosis mutant lacking the groEL homologue cpn60.1 is viable but fails to induce an inflammatory response in animal models of infection. *Infection and Immunity*, *76*(4), 1535–1546. https://doi.org/10.1128/IAI.01078-07

Hunter, R. L., Olsen, M., Jagannath, C., & Actor, J. K. (2006). Trehalose 6,6′-Dimycolate and Lipid in the Pathogenesis of Caseating Granulomas of Tuberculosis in Mice. *The American Journal of Pathology*, *168*(4), 1249–1261. https://doi.org/10.2353/ajpath.2006.050848

Jafari, M., & Ansari-Pour, N. (2019). Why, When and How to Adjust Your P Values? *Cell Journal*, *20*(4), 604–607. https://doi.org/10.22074/cellj.2019.5992

Jäger, S., Jäger, A., & Klug, G. (2004). CIRCE is not involved in heat-dependent transcription of groESL but in stabilization of the mRNA 5′-end in Rhodobacter capsulatus. *Nucleic Acids Research*, *32*(1), 386–396. https://doi.org/10.1093/nar/gkh174

Jain, S., Kaushal, D., DasGupta, S. K., & Tyagi, A. K. (1997). Construction of shuttle vectors for genetic manipulation and molecular analysis of mycobacteria. *Gene*, *190*(1), 37–44. https://doi.org/10.1016/s0378-1119(96)00746-9

Janowska, M. K., Baughman, H. E. R., Woods, C. N., & Klevit, R. E. (2019). Mechanisms of Small Heat Shock Proteins. *Cold Spring Harbor Perspectives in Biology*, *11*(10), a034025. https://doi.org/10.1101/cshperspect.a034025

Jeremiah, C., Petersen, E., Nantanda, R., Mungai, B. N., Migliori, G. B., Amanullah, F., Lungu, P., Ntoumi, F., Kumarasamy, N., Maeurer, M., & Zumla, A. (2022). The WHO Global Tuberculosis 2021 Report – not so good news and turning the tide back to End TB. *International Journal of Infectious Diseases*. https://doi.org/10.1016/j.ijid.2022.03.011

John, F. (2021). History of Tuberculosis. Part 1 – Phthisis, consumption and the White Plague. *Journal of Military and Veterans' Health.*, *Retrieved*.

Joseph, S., Yuen, A., Singh, V., & Hmama, Z. (2017). *Mycobacterium tuberculosis* Cpn60.2 (GroEL2) blocks macrophage apoptosis via interaction with mitochondrial mortalin. *Biology Open*. https://doi.org/10.1242/bio.023119

Kalate, R. N., Tambe, S. S., & Kulkarni, B. D. (2003). Artificial neural networks for prediction of mycobacterial promoter sequences. *Computational Biology and Chemistry*, *27*(6), 555–564. https://doi.org/10.1016/j.compbiolchem.2003.09.004

Kamal B. Sethna, N. F. M. (1998). Longitudinal Trends in Serum Levels of Mycobacterial Secretory (30 kD) and Cytoplasmic (65 kD) Antigens During Chemotherapy of Pulmonary Tuberculosis Patients. *Scandinavian Journal of Infectious Diseases*, *30*(4), 363–369. https://doi.org/10.1080/00365549850160657

Kanhere, A., & Bansal, M. (2005). A novel method for prokaryotic promoter prediction based on DNA stability. *BMC Bioinformatics*, *6*(1), 1. https://doi.org/10.1186/1471-2105-6-1

Kapopoulou, A., Lew, J. M., & Cole, S. T. (2011). The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis*, *91*(1), 8–13. https://doi.org/10.1016/j.tube.2010.09.006

Karunanayake, C., & Page, R. C. (2021). Cytosolic protein quality control machinery: Interactions of Hsp70 with a network of co-chaperones and substrates. *Experimental Biology and Medicine*, *246*(12), 1419–1434. https://doi.org/10.1177/1535370221999812

Kauffman, K. D., Sallin, M. A., Sakai, S., Kamenyeva, O., Kabat, J., Weiner, D., Sutphin, M., Schimel, D., Via, L., Barry, C. E., Wilder-Kofie, T., Moore, I., Moore, R., & Barber, D. L. (2018). Defective positioning in granulomas but not lung-homing limits CD4 T-cell interactions with Mycobacterium tuberculosis-infected macrophages in rhesus macaques. *Mucosal Immunology*, *11*(2), 462–473. https://doi.org/10.1038/mi.2017.60

Keller, P. J. (2013). In vivo imaging of zebrafish embryogenesis. *Methods*, *62*(3), 268–278. https://doi.org/10.1016/j.ymeth.2013.03.015

Khawbung, J. L., Nath, D., & Chakraborty, S. (2021). Drug resistant Tuberculosis: A review. *Comparative Immunology, Microbiology and Infectious Diseases*, *74*, 101574. https://doi.org/10.1016/j.cimid.2020.101574

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. https://doi.org/10.1186/gb-2013-14-4-r36

Kim, J., Lee, S., Shin, H., Kim, S. C., & Cho, B.-K. (2012). Elucidation of bacterial genome complexity using next-generation sequencing. *Biotechnology and Bioprocess Engineering*, *17*(5), 887–899. https://doi.org/10.1007/s12257-012-0374-x

Kim, Y. E., Hipp, M. S., Bracher, A., Hayer-Hartl, M., & Ulrich Hartl, F. (2013). Molecular Chaperone Functions in Protein Folding and Proteostasis. *Annual Review of Biochemistry*, *82*(1), 323–355. https://doi.org/10.1146/annurev-biochem-060208-092442

Kingston, W. (2004). Streptomycin, Schatz v. Waksman, and the Balance of Credit for Discovery. *Journal of the History of Medicine and Allied Sciences*, *59*(3), 441–462. https://doi.org/10.1093/jhmas/jrh091

Kinkar, E., Kinkar, A., & Saleh, M. (2019). The multicopper oxidase of Mycobacterium tuberculosis (MmcO) exhibits ferroxidase activity and scavenges reactive oxygen species in activated THP-1 cells. *International Journal of Medical Microbiology*, *309*(7), 151324. https://doi.org/10.1016/j.ijmm.2019.06.004

Kong, T. H., Coates, A. R. M., Butcher, P. D., Hickmant, C. J., & Shinnickt, T. M. (1993). Mycobacterium tuberculosis expresses two chaperonin-60 homologs. In *Proc. Natl. Acad. Sci. USA* (Vol. 90).

Koul, A., Arnoult, E., Lounis, N., Guillemont, J., & Andries, K. (2011). The challenge of new drug discovery for tuberculosis. *Nature*, *469*(7331), 483–490. https://doi.org/10.1038/nature09657

Kudryavtseva, S. S., Pichkur, E. B., Yaroshevich, I. A., Mamchur, A. A., Panina, I. S., Moiseenko, A. v., Sokolova, O. S., Muronetz, V. I., & Stanishneva-Konovalova, T. B. (2021). Novel cryo-EM structure of an ADP-bound GroEL–GroES complex. *Scientific Reports*, *11*(1), 18241. https://doi.org/10.1038/s41598-021-97657-x

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, *2015*(11), pdb.top084970. https://doi.org/10.1101/pdb.top084970

Kumar, C. M. S., Khare, G., Srikanth, C. v., Tyagi, A. K., Sardesai, A. A., & Mande, S. C. (2009a). Facilitated Oligomerization of Mycobacterial GroEL: Evidence for Phosphorylation-Mediated Oligomerization. *Journal of Bacteriology*, *191*(21), 6525–6538. https://doi.org/10.1128/JB.00652-09

Kumar, C. M. S., Khare, G., Srikanth, C. v., Tyagi, A. K., Sardesai, A. A., & Mande, S. C. (2009b). Facilitated Oligomerization of Mycobacterial GroEL: Evidence for Phosphorylation-Mediated Oligomerization. *Journal of Bacteriology*, *191*(21), 6525–6538. https://doi.org/10.1128/JB.00652-09

Kumar, C. M. S., Mande, S. C., & Mahajan, G. (2015). Multiple chaperonins in bacteria—novel functions and non-canonical behaviors. *Cell Stress and Chaperones*, *20*(4), 555–574. https://doi.org/10.1007/s12192-015-0598-8

Laale, H. W. (1977). The biology and use of zebrafish, Brachydanio rerio in fisheries research.. A literature review. *Journal of Fish Biology*, *10*(2), 121–173. https://doi.org/10.1111/j.1095-8649.1977.tb04049.x

Lai, H.-Y., Zhang, Z.-Y., Su, Z.-D., Su, W., Ding, H., Chen, W., & Lin, H. (2019). iProEP: A Computational Predictor for Predicting Promoter. *Molecular Therapy - Nucleic Acids*, *17*, 337–346. https://doi.org/10.1016/j.omtn.2019.05.028

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Laskey, R. A., Honda, B. M., Mills, A. D., & Finch, J. T. (1978). Nucleosomes are assembled by an acidic protein which binds histones and transfers them to DNA. *Nature*, *275*(5679), 416–420. https://doi.org/10.1038/275416a0

Lawson, N. D., & Wolfe, S. A. (2011). Forward and Reverse Genetic Approaches for the Analysis of Vertebrate Development in the Zebrafish. *Developmental Cell*, *21*(1), 48–64. https://doi.org/10.1016/j.devcel.2011.06.007

Lenaerts, A., Barry, C. E., & Dartois, V. (2015). Heterogeneity in tuberculosis pathology, microenvironments and therapeutic responses. *Immunological Reviews*, *264*(1), 288–307. https://doi.org/10.1111/imr.12252

Lenaerts, A. J., Hoff, D., Aly, S., Ehlers, S., Andries, K., Cantarero, L., Orme, I. M., & Basaraba, R. J. (2007). Location of Persisting Mycobacteria in a Guinea Pig Model of Tuberculosis Revealed by R207910. *Antimicrobial Agents and Chemotherapy*, *51*(9), 3338–3345. https://doi.org/10.1128/AAC.00276-07

Lewthwaite, J. C., Coates, A. R. M., Tormay, P., Singh, M., Mascagni, P., Poole, S., Roberts, M., Sharp, L., & Henderson, B. (2001). *Mycobacterium tuberculosis* Chaperonin 60.1 Is a More Potent Cytokine Stimulator than Chaperonin 60.2 (Hsp 65) and Contains a CD14-Binding Domain. *Infection and Immunity*, *69*(12), 7349–7355. https://doi.org/10.1128/IAI.69.12.7349-7355.2001

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323. https://doi.org/10.1186/1471-2105-12-323

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J., & Sha, B. (2003). Crystal Structure of the E. coli Hsp100 ClpB N-Terminal Domain. *Structure*, *11*(3), 323–328. https://doi.org/10.1016/S0969-2126(03)00030-3

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lieschke, G. J., & Currie, P. D. (2007). Animal models of human disease: zebrafish swim into view. *Nature Reviews Genetics*, *8*(5), 353–367. https://doi.org/10.1038/nrg2091

Lindquist, S., & Craig, E. A. (1988). THE HEAT-SHOCK PROTEINS. *Annual Review of Genetics*, *22*(1), 631–677. https://doi.org/10.1146/annurev.ge.22.120188.003215

Liu, J., Huang, C., Shin, D.-H., Yokota, H., Jancarik, J., Kim, J.-S., Adams, P. D., Kim, R., & Kim, S.-H. (2005). Crystal structure of a heat-inducible transcriptional repressor HrcA from Thermotoga maritima: structural insight into DNA binding and dimerization. *Journal of Molecular Biology*, *350*(5), 987–996. https://doi.org/10.1016/j.jmb.2005.04.021

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, *2012*, 1–11. https://doi.org/10.1155/2012/251364

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26. https://doi.org/10.1186/1748-7188-6-26

Lorenz, R., Wolfinger, M. T., Tanzer, A., & Hofacker, I. L. (2016). Predicting RNA secondary structures from sequence and probing data. *Methods (San Diego, Calif.)*, *103*, 86–98. https://doi.org/10.1016/j.ymeth.2016.04.004

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lund, P. A. (2001). *Microbial molecular chaperones* (pp. 93–140). https://doi.org/10.1016/S0065-2911(01)44012-4
Lund, P. A. (2009). Multiple chaperonins in bacteria - Why so many?: Review article. In *FEMS Microbiology Reviews* (Vol. 33, Issue 4, pp. 785–800). https://doi.org/10.1111/j.1574-6976.2009.00178.x
Luukinen, H., Hammarén, M. M., Vanha-aho, L.-M., & Parikka, M. (2018). Modeling Tuberculosis in Mycobacterium marinum Infected Adult Zebrafish. *Journal of Visualized Experiments*, *140*. https://doi.org/10.3791/58299
Maitra, A., Munshi, T., Healy, J., Martin, L. T., Vollmer, W., Keep, N. H., & Bhakta, S. (2019). Cell wall peptidoglycan in *Mycobacterium tuberculosis* : An Achilles' heel for the TB-causing pathogen. *FEMS Microbiology Reviews*, *43*(5), 548–575. https://doi.org/10.1093/femsre/fuz016
Malys, N. (2012). Shine-Dalgarno sequence of bacteriophage T4: GAGG prevails in early genes. *Molecular Biology Reports*, *39*(1), 33–39. https://doi.org/10.1007/s11033-011-0707-4
Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00426
Mardis, E. R. (2013). Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, *6*(1), 287–303. https://doi.org/10.1146/annurev-anchem-062012-092628
Marguerat, S., & Bähler, J. (2010). RNA-seq: from technology to biology. *Cellular and Molecular Life Sciences*, *67*(4), 569–579. https://doi.org/10.1007/s00018-009-0180-6
Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, *18*(9), 1509–1517. https://doi.org/10.1101/gr.079558.108
Matern, W. M., Jenquin, R. L., Bader, J. S., & Karakousis, P. C. (2020). Identifying the essential genes of Mycobacterium avium subsp. hominissuis with Tn-Seq using a rank-based filter procedure. *Scientific Reports*, *10*(1), 1095. https://doi.org/10.1038/s41598-020-57845-7
Mattila, J. T., Ojo, O. O., Kepka-Lenhart, D., Marino, S., Kim, J. H., Eum, S. Y., Via, L. E., Barry, C. E., Klein, E., Kirschner, D. E., Morris, S. M., Lin, P. L., & Flynn, J. L. (2013). Microenvironments in Tuberculous Granulomas Are Delineated by Distinct Populations of Macrophage Subsets and Expression of Nitric Oxide Synthase and Arginase Isoforms. *The Journal of Immunology*, *191*(2), 773–784. https://doi.org/10.4049/jimmunol.1300113
McDermaid, A., Monier, B., Zhao, J., Liu, B., & Ma, Q. (2019). Interpretation of differential gene expression results of RNA-seq data: review and integration. *Briefings in Bioinformatics*, *20*(6), 2044–2054. https://doi.org/10.1093/bib/bby067
McLennan, N., & Masters, M. (1998). GroE is vital for cell-wall synthesis. *Nature*, *392*(6672), 139–139. https://doi.org/10.1038/32317
Meghji, S., White, P. A., Nair, S. P., Reddi, K., Heron, K., Henderson, B., Zaliani, A., Fossati, G., Mascagni, P., Hunt, J. F., Roberts, M. M., & Coates, A. R. M. (1997). Mycobacterium tuberculosis Chaperonin 10 Stimulates Bone Resorption: A

Potential Contributory Factor in Pott's Disease. *Journal of Experimental Medicine*, *186*(8), 1241–1246. https://doi.org/10.1084/jem.186.8.1241
Melly, G., & Purdy, G. (2019). MmpL Proteins in Physiology and Pathogenesis of M. tuberculosis. *Microorganisms*, *7*(3), 70. https://doi.org/10.3390/microorganisms7030070
Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626
Minder, A. C., Fischer, H.-M., Hennecke, H., & Narberhaus, F. (2000). Role of HrcA and CIRCE in the Heat Shock Regulatory Network of Bradyrhizobium japonicum. In *JOURNAL OF BACTERIOLOGY* (Vol. 182, Issue 1). https://journals.asm.org/journal/jb
Mishra, K. C., de Chastellier, C., Narayana, Y., Bifani, P., Brown, A. K., Besra, G. S., Katoch, V. M., Joshi, B., Balaji, K. N., & Kremer, L. (2008). Functional Role of the PE Domain and Immunogenicity of the *Mycobacterium tuberculosis* Triacylglycerol Hydrolase LipY. *Infection and Immunity*, *76*(1), 127–140. https://doi.org/10.1128/IAI.00410-07
Mogk, A. (1997). The GroE chaperonin machine is a major modulator of the CIRCE heat shock regulon of Bacillus subtilis. *The EMBO Journal*, *16*(15), 4579–4590. https://doi.org/10.1093/emboj/16.15.4579
Moody, D. E. (2001). Genomics techniques: An overview of methods for the study of gene expression. *Journal of Animal Science*, *79*(E-Suppl), E128. https://doi.org/10.2527/jas2001.79E-SupplE128x
Munjal, G., Hanmandlu, M., & Srivastava, S. (2019). *Phylogenetics Algorithms and Applications* (pp. 187–194). https://doi.org/10.1007/978-981-13-5934-7_17
Munshi, T., Gupta, A., Evangelopoulos, D., Guzman, J. D., Gibbons, S., Keep, N. H., & Bhakta, S. (2013). Characterisation of ATP-Dependent Mur Ligases Involved in the Biogenesis of Cell Wall Peptidoglycan in Mycobacterium tuberculosis. *PLoS ONE*, *8*(3), e60143. https://doi.org/10.1371/journal.pone.0060143
Myllymäki, H., Bäuerlein, C. A., & Rämet, M. (2016). The Zebrafish Breathes New Life into the Study of Tuberculosis. *Frontiers in Immunology*, *7*. https://doi.org/10.3389/fimmu.2016.00196
Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, *320*(5881), 1344–1349. https://doi.org/10.1126/science.1158441
Narayan, A., Sachdeva, P., Sharma, K., Saini, A. K., Tyagi, A. K., & Singh, Y. (2007). Serine threonine protein kinases of mycobacterial genus: phylogeny to function. *Physiological Genomics*, *29*(1), 66–75. https://doi.org/10.1152/physiolgenomics.00221.2006
Narberhaus, F. (1999). Negative regulation of bacterial heat shock genes. *Molecular Microbiology*, *31*(1), 1–8. https://doi.org/10.1046/j.1365-2958.1999.01166.x
Narberhaus, F., & Bahl, H. (1992). Cloning, sequencing, and molecular analysis of the groESL operon of Clostridium acetobutylicum. *Journal of Bacteriology*, *174*(10), 3282–3289. https://doi.org/10.1128/jb.174.10.3282-3289.1992
Nelson Onuchic, J., Nymeyer, H., García, A. E., Chahine, J., & Socci, N. D. (2000). *The energy landscape theory of protein folding: Insights into folding mechanisms and scenarios* (pp. 87–152). https://doi.org/10.1016/S0065-3233(00)53003-4
Nielsen, K. L., McLennan, N., Masters, M., & Cowan, N. J. (1999). A Single-Ring Mitochondrial Chaperonin (Hsp60-Hsp10) Can Substitute for GroEL-GroES In

Vivo. *Journal of Bacteriology*, *181*(18), 5871–5875. https://doi.org/10.1128/JB.181.18.5871-5875.1999
Ojha, A., Anand, M., Bhatt, A., Kremer, L., Jacobs, W. R., & Hatfull, G. F. (2005). GroEL1: A dedicated chaperone involved in mycolic acid biosynthesis during biofilm formation in mycobacteria. *Cell*, *123*(5), 861–873. https://doi.org/10.1016/j.cell.2005.09.012
Oubounyt, M., Louadi, Z., Tayara, H., & Chong, K. T. (2019). DeePromoter: Robust Promoter Predictor Using Deep Learning. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00286
Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421
Paju, S., Goulhen, F., Asikainen, S., Grenier, D., Mayrand, D., & Uitto, V.-J. (2000). Localization of heat shock proteins in clinical *Actinobacillus actinomycetemcomitans* strains and their effects on epithelial cell proliferation. *FEMS Microbiology Letters*, *182*(2), 231–235. https://doi.org/10.1111/j.1574-6968.2000.tb08900.x
Papadimitriou, K., Alegría, Á., Bron, P. A., de Angelis, M., Gobbetti, M., Kleerebezem, M., Lemos, J. A., Linares, D. M., Ross, P., Stanton, C., Turroni, F., van Sinderen, D., Varmanen, P., Ventura, M., Zúñiga, M., Tsakalidou, E., & Kok, J. (2016). Stress Physiology of Lactic Acid Bacteria. *Microbiology and Molecular Biology Reviews*, *80*(3), 837–890. https://doi.org/10.1128/MMBR.00076-15
Parijat, P., & Batra, J. K. (2015). Role of DnaK in HspR-HAIR interaction of *Mycobacterium tuberculosis*. *IUBMB Life*, *67*(11), 816–827. https://doi.org/10.1002/iub.1438
Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE*, *7*(2), e30619. https://doi.org/10.1371/journal.pone.0030619
Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197
Pease, J., & Sooknanan, R. (2012). A rapid, directional RNA-seq library preparation workflow for Illumina® sequencing. *Nature Methods*, *9*(3), i–ii. https://doi.org/10.1038/nmeth.f.355
Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, *11*(9), 1650–1667. https://doi.org/10.1038/nprot.2016.095
Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122
Pinto, A. C., Melo-Barbosa, H. P., Miyoshi, A., Silva, A., & Azevedo, V. (2011). Review Application of RNA-seq to reveal the transcript profile in bacteria. *Genetics and Molecular Research*, *10*(3), 1707–1718. https://doi.org/10.4238/vol10-3gmr1554
Poulsen, L. D., & Vinther, J. (2018). RNA-Seq for Bacterial Gene Expression. *Current Protocols in Nucleic Acid Chemistry*, *73*(1), e55. https://doi.org/10.1002/cpnc.55

Prasanna, A. N., & Mehra, S. (2013). Comprehensive Phylogenetic Analysis of Mycobacteria. *IFAC Proceedings Volumes*, *46*(31), 101–106. https://doi.org/10.3182/20131216-3-IN-2044.00050

Puigbò, P., Bravo, I. G., & Garcia-Vallve, S. (2008). CAIcal: A combined set of tools to assess codon usage adaptation. *Biology Direct*, *3*(1), 38. https://doi.org/10.1186/1745-6150-3-38

Qamra, R., & Mande, S. C. (2004). Crystal Structure of the 65-Kilodalton Heat Shock Protein, Chaperonin 60.2, of *Mycobacterium tuberculosis*. *Journal of Bacteriology*, *186*(23), 8105–8113. https://doi.org/10.1128/JB.186.23.8105-8113.2004

Qamra, R., Srinivas, V., & Mande, S. C. (2004). Mycobacterium tuberculosis GroEL Homologues Unusually Exist as Lower Oligomers and Retain the Ability to Suppress Aggregation of Substrate Proteins. *Journal of Molecular Biology*, *342*(2), 605–617. https://doi.org/10.1016/j.jmb.2004.07.066

Quadri, L. E. N. (2014). Biosynthesis of mycobacterial lipids by polyketide synthases and beyond. *Critical Reviews in Biochemistry and Molecular Biology*, *49*(3), 179–211. https://doi.org/10.3109/10409238.2014.896859

Quail, M., Smith, M. E., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, *13*(1), 341. https://doi.org/10.1186/1471-2164-13-341

Quax, T. E. F., Claassens, N. J., Söll, D., & van der Oost, J. (2015). Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell*, *59*(2), 149–161. https://doi.org/10.1016/j.molcel.2015.05.035

Ramakrishnan, L. (2012). Revisiting the role of the granuloma in tuberculosis. *Nature Reviews Immunology*, *12*(5), 352–366. https://doi.org/10.1038/nri3211

Raman, S., Hazra, R., Dascher, C. C., & Husson, R. N. (2004). Transcription Regulation by the *Mycobacterium tuberculosis* Alternative Sigma Factor SigD and Its Role in Virulence. *Journal of Bacteriology*, *186*(19), 6605–6616. https://doi.org/10.1128/JB.186.19.6605-6616.2004

Ranford, J. C., Coates, A. R. M., & Henderson, B. (2000). Chaperonins are cell-signalling proteins: the unfolding biology of molecular chaperones. *Expert Reviews in Molecular Medicine*, *2*(8), 1–17. https://doi.org/10.1017/S1462399400002015

Ranson, N. A., Clare, D. K., Farr, G. W., Houldershaw, D., Horwich, A. L., & Saibil, H. R. (2006). Allosteric signaling of ATP hydrolysis in GroEL–GroES complexes. *Nature Structural & Molecular Biology*, *13*(2), 147–152. https://doi.org/10.1038/nsmb1046

Rao, M., Ippolito, G., Mfinanga, S., Ntoumi, F., Yeboah-Manu, D., Vilaplana, C., Zumla, A., & Maeurer, M. (2019). Latent TB Infection (LTBI) – Mycobacterium tuberculosis pathogenesis and the dynamics of the granuloma battleground. *International Journal of Infectious Diseases*, *80*, S58–S61. https://doi.org/10.1016/j.ijid.2019.02.035

Rao, T., & Lund, P. A. (2010). Differential expression of the multiple chaperonins of Mycobacterium smegmatis. *FEMS Microbiology Letters*, *310*(1), 24–31. https://doi.org/10.1111/j.1574-6968.2010.02039.x

Rao, Tara. (2010). *Analysis of the multiple chaperonins of Mycobacterium smegmatis.* University of Birmingham .

Richter, K., Haslbeck, M., & Buchner, J. (2010). The Heat Shock Response: Life on the Verge of Death. *Molecular Cell*, *40*(2), 253–266. https://doi.org/10.1016/j.molcel.2010.10.006

Rifat, D., Chen, L., Kreiswirth, B. N., & Nuermberger, E. L. (2021). Genome-Wide Essentiality Analysis of *Mycobacterium abscessus* by Saturated Transposon Mutagenesis and Deep Sequencing. *MBio*, *12*(3). https://doi.org/10.1128/mBio.01049-21

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, M. D., & Oshlack, A. (2010a). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Robinson, M. D., & Oshlack, A. (2010b). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Roncarati, D., & Scarlato, V. (2017). Regulation of heat-shock genes in bacteria: from signal sensing to gene expression output. *FEMS Microbiology Reviews*, *41*(4), 549–574. https://doi.org/10.1093/femsre/fux015

Rosano, G. L., & Ceccarelli, E. A. (2009). Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted Escherichia coli strain. *Microbial Cell Factories*, *8*(1), 41. https://doi.org/10.1186/1475-2859-8-41

Röse, L., Kaufmann, S. H. E., & Daugelat, S. (2004). Involvement of Mycobacterium smegmatis undecaprenyl phosphokinase in biofilm and smegma formation. *Microbes and Infection*, *6*(11), 965–971. https://doi.org/10.1016/j.micinf.2004.05.011

Roy, S., Ghatak, D., Das, P., & BoseDasgupta, S. (2020). ESX secretion system: The gatekeepers of mycobacterial survivability and pathogenesis. *European Journal of Microbiology and Immunology*, *10*(4), 202–209. https://doi.org/10.1556/1886.2020.00028

Russell, D. G., Cardona, P.-J., Kim, M.-J., Allain, S., & Altare, F. (2009). Foamy macrophages and the progression of the human tuberculosis granuloma. *Nature Immunology*, *10*(9), 943–948. https://doi.org/10.1038/ni.1781

Rye, H. S., Burston, S. G., Fenton, W. A., Beechem, J. M., Xu, Z., Sigler, P. B., & Horwich, A. L. (1997). Distinct actions of cis and trans ATP within the double ring of the chaperonin GroEL. *Nature*, *388*(6644), 792–798. https://doi.org/10.1038/42047

Sabin, F. R. (1932). CELLULAR REACTIONS TO FRACTIONS ISOLATED FROM TUBERCLE BACILLI. *Physiological Reviews*, *12*(2), 141–165. https://doi.org/10.1152/physrev.1932.12.2.141

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sato, S., Ikeuchi, M., & Nakamoto, H. (2008). Expression and function of a groEL paralog in the thermophilic cyanobacterium Thermosynechococcus elongatus

under heat and cold stress. *FEBS Letters*, *582*(23–24), 3389–3395. https://doi.org/10.1016/j.febslet.2008.08.034
Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863–864. https://doi.org/10.1093/bioinformatics/btr026
Schön, U., & Schumann, W. (1993). Molecular cloning, sequencing, and transcriptional analysis of the groESL operon from Bacillus stearothermophilus. *Journal of Bacteriology*, *175*(8), 2465–2469. https://doi.org/10.1128/jb.175.8.2465-2469.1993
Senkovich, O., Ceaser, S., McGee, D. J., & Testerman, T. L. (2010). Unique Host Iron Utilization Mechanisms of *Helicobacter pylori* Revealed with Iron-Deficient Chemically Defined Media. *Infection and Immunity*, *78*(5), 1841–1849. https://doi.org/10.1128/IAI.01258-09
Servant, P., & Mazodier, P. (2001). Negative regulation of the heat shock response in Streptomyces. In *Archives of Microbiology* (Vol. 176, Issue 4, pp. 237–242). https://doi.org/10.1007/s002030100321
Seyednasrollah, F., Laiho, A., & Elo, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, *16*(1), 59–70. https://doi.org/10.1093/bib/bbt086
Sharp, P. M. (2005). Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research*, *33*(4), 1141–1153. https://doi.org/10.1093/nar/gki242
Shell, S. S., Wang, J., Lapierre, P., Mir, M., Chase, M. R., Pyle, M. M., Gawande, R., Ahmad, R., Sarracino, D. A., Ioerger, T. R., Fortune, S. M., Derbyshire, K. M., Wade, J. T., & Gray, T. A. (2015a). Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS Genetics*, *11*(11), e1005641. https://doi.org/10.1371/journal.pgen.1005641
Shell, S. S., Wang, J., Lapierre, P., Mir, M., Chase, M. R., Pyle, M. M., Gawande, R., Ahmad, R., Sarracino, D. A., Ioerger, T. R., Fortune, S. M., Derbyshire, K. M., Wade, J. T., & Gray, T. A. (2015b). Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS Genetics*, *11*(11). https://doi.org/10.1371/journal.pgen.1005641
Shim, D., Kim, H., & Shin, S. J. (2020). Mycobacterium tuberculosis Infection-Driven Foamy Macrophages and Their Implications in Tuberculosis Control as Targets for Host-Directed Therapy. *Frontiers in Immunology*, *11*. https://doi.org/10.3389/fimmu.2020.00910
Shinnick, T. M., Vodkin, M. H., & Williams, J. C. (1988). The Mycobacterium tuberculosis 65-kilodalton antigen is a heat shock protein which corresponds to common antigen and to the Escherichia coli GroEL protein. *Infection and Immunity*, *56*(2), 446–451. https://doi.org/10.1128/iai.56.2.446-451.1988
Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., & Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, *100*(26), 15776–15781. https://doi.org/10.1073/pnas.2136655100
Sielaff, B., Lee, K. S., & Tsai, F. T. F. (2011). Structural and Functional Conservation of Mycobacterium tuberculosis GroEL Paralogs Suggests that GroEL1 Is a

Chaperonin. *Journal of Molecular Biology*, *405*(3), 831–839. https://doi.org/10.1016/j.jmb.2010.11.021
Silva, C. L., Ekizlerian, S. M., & Fazioli, R. A. (1985). Role of cord factor in the modulation of infection caused by mycobacteria. *The American Journal of Pathology*, *118*(2), 238–247.
Silva Miranda, M., Breiman, A., Allain, S., Deknuydt, F., & Altare, F. (2012). The Tuberculous Granuloma: An Unsuccessful Host Defence Mechanism Providing a Safety Shelter for the Bacteria? *Clinical and Developmental Immunology*, *2012*, 1–14. https://doi.org/10.1155/2012/139127
Silva, S. de A. e, Gerhardt, G. J. L., & Echeverrigaray, S. (2011). Rules extraction from neural networks applied to the prediction and recognition of prokaryotic promoters. *Genetics and Molecular Biology*, *34*(2), 353–360. https://doi.org/10.1590/S1415-47572011000200031
Siméone, R., Léger, M., Constant, P., Malaga, W., Marrakchi, H., Daffé, M., Guilhot, C., & Chalut, C. (2010). Delineation of the roles of FadD22, FadD26 and FadD29 in the biosynthesis of phthiocerol dimycocerosates and related compounds in Mycobacterium tuberculosis. *FEBS Journal*, *277*(12), 2715–2725. https://doi.org/10.1111/j.1742-4658.2010.07688.x
Sinigaglia, A., Peta, E., Riccetti, S., Venkateswaran, S., Manganelli, R., & Barzon, L. (2020). Tuberculosis-Associated MicroRNAs: From Pathogenesis to Disease Biomarkers. *Cells*, *9*(10), 2160. https://doi.org/10.3390/cells9102160
Sivinski, J., Ambrose, A. J., Panfilenko, I., Zerio, C. J., Machulis, J. M., Mollasalehi, N., Kaneko, L. K., Stevens, M., Ray, A.-M., Park, Y., Wu, C., Hoang, Q. Q., Johnson, S. M., & Chapman, E. (2021). Functional Differences between E. coli and ESKAPE Pathogen GroES/GroEL. *MBio*, *12*(1). https://doi.org/10.1128/mBio.02167-20
Skjærven, L., Reuter, N., & Martinez, A. (2011). Dynamics, flexibility and ligand-induced conformational changes in biological macromolecules: a computational approach. *Future Medicinal Chemistry*, *3*(16), 2079–2100. https://doi.org/10.4155/fmc.11.159
Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1). https://doi.org/10.1002/cpmb.59
Smale, S. T. (2010). β-Galactosidase Assay. *Cold Spring Harbor Protocols*, *2010*(5), pdb.prot5423. https://doi.org/10.1101/pdb.prot5423
Smid, M., Coebergh van den Braak, R. R. J., van de Werken, H. J. G., van Riet, J., van Galen, A., de Weerd, V., van der Vlugt-Daane, M., Bril, S. I., Lalmahomed, Z. S., Kloosterman, W. P., Wilting, S. M., Foekens, J. A., IJzermans, J. N. M., Martens, J. W. M., & Sieuwerts, A. M. (2018). Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons. *BMC Bioinformatics*, *19*(1), 236. https://doi.org/10.1186/s12859-018-2246-7
Sorek, R., & Cossart, P. (2010). Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature Reviews Genetics*, *11*(1), 9–16. https://doi.org/10.1038/nrg2695
Srinivasan, K. A., Virdee, S. K., & McArthur, A. G. (2020). Strandedness during cDNA synthesis, the stranded parameter in htseq-count and analysis of RNA-Seq data. *Briefings in Functional Genomics*, *19*(5–6), 339–342. https://doi.org/10.1093/bfgp/elaa010

Stewart, G. R., Snewin, V. A., Walzl, G., Hussell, T., Tormay, P., O'Gaora, P., Goyal, M., Betts, J., Brown, I. N., & Young, D. B. (2001). Overexpression of heat-shock proteins reduces survival of Mycobacterium tuberculosis in the chronic phase of infection. *Nature Medicine*, *7*(6), 732–737. https://doi.org/10.1038/89113

Stewart, G. R., Wernisch, L., Stabler, R., Mangan, J. A., Hinds, J., Laing, K. G., Young, D. B., Butcher, P. D., & Stewart, G. (2002). Dissection of the heat-shock response in Mycobacterium tuberculosis using mutants and microarrays. In *Microbiology* (Vol. 148).

Stinear, T. P., Seemann, T., Harrison, P. F., Jenkin, G. A., Davies, J. K., Johnson, P. D. R., Abdellah, Z., Arrowsmith, C., Chillingworth, T., Churcher, C., Clarke, K., Cronin, A., Davis, P., Goodhead, I., Holroyd, N., Jagels, K., Lord, A., Moule, S., Mungall, K., … Cole, S. T. (2008). Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*. *Genome Research*, *18*(5), 729–741. https://doi.org/10.1101/gr.075069.107

Stock, J. B., Ninfa, A. J., & Stock, A. M. (1989a). Protein phosphorylation and regulation of adaptive responses in bacteria. *Microbiological Reviews*, *53*(4), 450–490. https://doi.org/10.1128/mr.53.4.450-490.1989

Stock, J. B., Ninfa, A. J., & Stock, A. M. (1989b). Protein phosphorylation and regulation of adaptive responses in bacteria. *Microbiological Reviews*, *53*(4), 450–490. https://doi.org/10.1128/mr.53.4.450-490.1989

Styblo, K. (1990). The global aspects of tuberculosis and HIV infection. *Bulletin of the International Union against Tuberculosis and Lung Disease*, *65*(1), 28–32.

Sudhagar, A., Kumar, G., & El-Matbouli, M. (2018). Transcriptome Analysis Based on RNA-Seq in Understanding Pathogenic Mechanisms of Diseases and the Immune System of Fish: A Comprehensive Review. *International Journal of Molecular Sciences*, *19*(1), 245. https://doi.org/10.3390/ijms19010245

Tegel, H., Ottosson, J., & Hober, S. (2011). Enhancing the protein production levels in Escherichia coli with a strong promoter. *FEBS Journal*, *278*(5), 729–739. https://doi.org/10.1111/j.1742-4658.2010.07991.x

Teng, R., & Dick, T. (2003). Isoniazid resistance of exponentially growing *Mycobacterium smegmatis* biofilm culture. *FEMS Microbiology Letters*, *227*(2), 171–174. https://doi.org/10.1016/S0378-1097(03)00584-6

Terpe, K. (2006). Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. *Applied Microbiology and Biotechnology*, *72*(2), 211–222. https://doi.org/10.1007/s00253-006-0465-8

Thulasiraman, V. (1999). In vivo newly translated polypeptides are sequestered in a protected folding environment. *The EMBO Journal*, *18*(1), 85–95. https://doi.org/10.1093/emboj/18.1.85

Tobin, D. M., & Ramakrishnan, L. (2008). Comparative pathogenesis of Mycobacterium marinum and Mycobacterium tuberculosis. *Cellular Microbiology*, *10*(5), 1027–1039. https://doi.org/10.1111/j.1462-5822.2008.01133.x

Tønjum, T., Welty, D. B., Jantzen, E., & Small, P. L. (1998). Differentiation of *Mycobacterium ulcerans* , *M. marinum* , and *M. haemophilum* : Mapping of Their Relationships to *M. tuberculosis* by Fatty Acid Profile Analysis, DNA-DNA Hybridization, and 16S rRNA Gene Sequence Analysis. *Journal of Clinical Microbiology*, *36*(4), 918–925. https://doi.org/10.1128/JCM.36.4.918-925.1998

Torraca, V., & Mostowy, S. (2018). Zebrafish Infection: From Pathogenesis to Cell Biology. *Trends in Cell Biology*, *28*(2), 143–156. https://doi.org/10.1016/j.tcb.2017.10.002

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–578. https://doi.org/10.1038/nprot.2012.016

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

Valenzuela-Valderas, K. N., Moreno-Hagelsieb, G., Rohde, J. R., & Garduño, R. A. (2021). The Functional Differences between the GroEL Chaperonin of Escherichia coli and the HtpB Chaperonin of Legionella pneumophila Can Be Mapped to Specific Amino Acid Residues. *Biomolecules*, *12*(1), 59. https://doi.org/10.3390/biom12010059

van Eden, W., van der Zee, R., Paul, A. G. A., Prakken, B. J., Wendling, U., Anderton, S. M., & Wauben, M. H. M. (1998). Do heat shock proteins control the balance of T-cell regulation in inlammatory diseases? *Immunology Today*, *19*(7), 303–307. https://doi.org/10.1016/S0167-5699(98)01283-3

van Vliet, A. H. M. (2010). Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiology Letters*, *302*(1), 1–7. https://doi.org/10.1111/j.1574-6968.2009.01767.x

Vanet, A., & Labigne, A. (1998). Evidence for Specific Secretion Rather than Autolysis in the Release of Some *Helicobacter pylori* Proteins. *Infection and Immunity*, *66*(3), 1023–1027. https://doi.org/10.1128/IAI.66.3.1023-1027.1998

Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial Analysis of Gene Expression. *Science*, *270*(5235), 484–487. https://doi.org/10.1126/science.270.5235.484

Vilchèze, C., & Jacobs, W. R. (2019). The Isoniazid Paradigm of Killing, Resistance, and Persistence in Mycobacterium tuberculosis. *Journal of Molecular Biology*, *431*(18), 3450–3461. https://doi.org/10.1016/j.jmb.2019.02.016

Vinod, V., Vijayrajratnam, S., Vasudevan, A. K., & Biswas, R. (2020). The cell surface adhesins of Mycobacterium tuberculosis. *Microbiological Research*, *232*, 126392. https://doi.org/10.1016/j.micres.2019.126392

Volkman, H. E., Clay, H., Beery, D., Chang, J. C. W., Sherman, D. R., & Ramakrishnan, L. (2004). Tuberculous Granuloma Formation Is Enhanced by a Mycobacterium Virulence Determinant. *PLoS Biology*, *2*(11), e367. https://doi.org/10.1371/journal.pbio.0020367

Wagner, G. P., Kin, K., & Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, *131*(4), 281–285. https://doi.org/10.1007/s12064-012-0162-3

Wang, Y., & DeHaseth, P. L. (2003). Sigma 32-dependent promoter activity in vivo: Sequence determinants of the groE promoter. *Journal of Bacteriology*, *185*(19), 5800–5806. https://doi.org/10.1128/JB.185.19.5800-5806.2003

Wehenkel, A., Bellinzoni, M., Graña, M., Duran, R., Villarino, A., Fernandez, P., Andre-Leroux, G., England, P., Takiff, H., Cerveñansky, C., Cole, S. T., & Alzari, P. M. (2008). Mycobacterial Ser/Thr protein kinases and phosphatases: Physiological

roles and therapeutic potential. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1784*(1), 193–202. https://doi.org/10.1016/j.bbapap.2007.08.006

Weissman, J. S., Rye, H. S., Fenton, W. A., Beechem, J. M., & Horwich, A. L. (1996). Characterization of the Active Intermediate of a GroEL–GroES-Mediated Protein Folding Reaction. *Cell*, *84*(3), 481–490. https://doi.org/10.1016/S0092-8674(00)81293-3

Wen, J.-D., Lancaster, L., Hodges, C., Zeri, A.-C., Yoshimura, S. H., Noller, H. F., Bustamante, C., & Tinoco, I. (2008). Following translation by single ribosomes one codon at a time. *Nature*, *452*(7187), 598–603. https://doi.org/10.1038/nature06716

Wetzstein, M., Volker, U., Dedio,', J., Lobau,', S., Zuber, U., Schiesswohl,' Christian Herget,', M., Hecker, M., & Schumann, W. (1992). Cloning, Sequencing, and Molecular Analysis of the dnaK Locus from Bacillus subtilis. In *JOURNAL OF BACTERIOLOGY* (Vol. 174, Issue 10). https://journals.asm.org/journal/jb

Wiegert, T., & Schumann, W. (2003). Analysis of a DNA-binding motif of the Bacillus subtilis HrcA repressor protein. *FEMS Microbiology Letters*, *223*(1), 101–106. https://doi.org/10.1016/S0378-1097(03)00350-1

Wolynes, P. G., Onuchic, J. N., & Thirumalai, D. (1995). Navigating the Folding Routes. *Science*, *267*(5204), 1619–1620. https://doi.org/10.1126/science.7886447

Wong, S.-L. (1995). Isolation and Characterization of Bacillus subtilis groE Regulatory Mutants: Evidence for orf39 in the dnaK Operon as a Repressor Gene in Regulating the Expression of Both groE and dnaK. In *JOURNAL OF BACTERIOLOGY* (Vol. 177, Issue 22). https://journals.asm.org/journal/jb

World Health Organization (WHO). (2021). *GLOBAL TUBERCULOSIS REPORT 2021*. http://apps.who.int/bookorders.

Xu, Z., Horwich, A. L., & Sigler, P. B. (1997a). The crystal structure of the asymmetric GroEL–GroES–(ADP)7 chaperonin complex. *Nature*, *388*(6644), 741–750. https://doi.org/10.1038/41944

Xu, Z., Horwich, A. L., & Sigler, P. B. (1997b). The crystal structure of the asymmetric GroEL–GroES–(ADP)7 chaperonin complex. *Nature*, *388*(6644), 741–750. https://doi.org/10.1038/41944

Xu, Z., & Sigler, P. B. (1998). GroEL/GroES: Structure and Function of a Two-Stroke Folding Machine. *Journal of Structural Biology*, *124*(2–3), 129–141. https://doi.org/10.1006/jsbi.1998.4060

Yip, M. J., Porter, J. L., Fyfe, J. A. M., Lavender, C. J., Portaels, F., Rhodes, M., Kator, H., Colorni, A., Jenkin, G. A., & Stinear, T. (2007). Evolution of *Mycobacterium ulcerans* and Other Mycolactone-Producing Mycobacteria from a Common *Mycobacterium marinum* Progenitor. *Journal of Bacteriology*, *189*(5), 2021–2029. https://doi.org/10.1128/JB.01442-06

Yoshida, M., Fukano, H., Miyamoto, Y., Shibayama, K., Suzuki, M., & Hoshino, Y. (2018). Complete Genome Sequence of Mycobacterium marinum ATCC 927 ^{T} , Obtained Using Nanopore and Illumina Sequencing Technologies. *Genome Announcements*, *6*(20). https://doi.org/10.1128/genomeA.00397-18

Yoshida, N., Oeda, K., Watanabe, E., Mikami, T., Fukita, Y., Nishimura, K., Komai, K., & Matsuda, K. (2001). Chaperonin turned insect toxin. *Nature*, *411*(6833), 44–44. https://doi.org/10.1038/35075148

Yuan, G., & Wong, S. L. (1995). Isolation and characterization of Bacillus subtilis groE regulatory mutants: evidence for orf39 in the dnaK operon as a repressor gene in

regulating the expression of both groE and dnaK. *Journal of Bacteriology*, *177*(22), 6462–6468. https://doi.org/10.1128/jb.177.22.6462-6468.1995
Zambrano, M. M., & Kolter, R. (2005). Mycobacterial Biofilms: A Greasy Way to Hold It Together. *Cell*, *123*(5), 762–764. https://doi.org/10.1016/j.cell.2005.11.011
Zeng, S., Constant, P., Yang, D., Baulard, A., Lefèvre, P., Daffé, M., Wattiez, R., & Fontaine, V. (2019). Cpn60.1 (GroEL1) contributes to mycobacterial crabtree effect: Implications for biofilm formation. *Frontiers in Microbiology*, *10*(JUN). https://doi.org/10.3389/fmicb.2019.01149
Zhao, S., Ye, Z., & Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA*, *26*(8), 903–909. https://doi.org/10.1261/rna.074922.120
Zhao, Y., Li, M.-C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., Williams, P. M., Evrard, Y. A., Doroshow, J. H., & McShane, L. M. (2021). TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine*, *19*(1), 269. https://doi.org/10.1186/s12967-021-02936-w
Zheng, H., He, J., Li, J., Yang, J., Kirk, M. L., Roman, L. J., & Feng, C. (2019). Generation and characterization of functional phosphoserine-incorporated neuronal nitric oxide synthase holoenzyme. *JBIC Journal of Biological Inorganic Chemistry*, *24*(1), 1–9. https://doi.org/10.1007/s00775-018-1621-1
Zuber, U., & Schumann, W. (1994). *CIRCE, a Novel Heat Shock Element Involved in Regulation of Heat Shock Operon dnaK of Bacillus subtilis* (Vol. 176, Issue 5). https://journals.asm.org/journal/jb
Zürcher, K., Zwahlen, M., Ballif, M., Rieder, H. L., Egger, M., & Fenner, L. (2016). Influenza Pandemics and Tuberculosis Mortality in 1889 and 1918: Analysis of Historical Data from Switzerland. *PLOS ONE*, *11*(10), e0162575. https://doi.org/10.1371/journal.pone.0162575